NLP and IR Research Group

ETSI Informática

Universidad Nacional de Educación a Distancia (UNED)

C/Juan del Rosal 16, 28040 Madrid (Spain)

A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet

Technical Report TR-2016-01

Juan J. Lastra-Díaz[1] Ana García-Serrano[2]

July 6, 2016

Cite this work as:

Lastra-Díaz, J. J., and García-Serrano, A. (2016). A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet. Technical Report TR-2016-01. NLP and IR Research Group. ETSI Informática. Universidad Nacional de Educación a Distancia (UNED). http://e-spacio.uned.es/fez/view/bibliuned:DptoLSI-ETSI-Informes-Jlastra-refinement

[1] jlastra@invi.uned.es (corresponding author)
[2] agarcia@lsi.uned.es

This page is intentionally left in blank

A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet

Juan J. Lastra-Díaz Ana García-Serrano
(jlastra@invi.uned.es, agarcia@lsi.uned.es)

NLP and IR Research Group
ETSI Informática
Universidad Nacional de Educación a Distancia (UNED)
C/Juan del Rosal 16, 28040 Madrid (Spain)

July 11, 2016

Abstract

In a recent paper, we introduce a new family of Information Content (IC) models based on the estimation of the conditional probability between child and parent concepts. This work is encouraged by the finding of two drawbacks in the computational method of our aforementioned family of IC models, as well as other two gaps in the literature. First gap is that two of our cognitive IC models do not satisfy the axiom that constrains the sum of probabilities on the leaf nodes to be 1, whilst some ontologies with multiple inheritance could prevent the IC model satisfying the *growing monotonicity axiom* in concepts with multiple parents. Second gap is the lack of a complete and updated experimental survey including a pairwise statistical significance analysis between most IC models and ontology-based similarity measures. Finally a third gap is the lack of replication and confirmation of previous methods and results in most works. The latest two gaps are especially significant in the current state of the problem, in which there is no convincing winner within the family of intrinsic IC-based similarity measures and the performance margin is very narrow. In order to bridge the aforementioned gaps, this paper introduces the following contributions: (1) a refinement of our recent family of well-founded Information Content (IC) models; (2) eight new intrinsic IC models and one new corpus-based IC model; and (3) a very detailed experimental survey of ontology-based similarity measures and Information Content (IC) models on WordNet, including the evaluation and statistical significance analysis on the five most significant datasets of most ontology-based similarity measures and all WordNet-based IC models reported in the literature, with the only exception of the IC models recently introduced by Harispe et al. (2015a) and Ben Aouicha et al. (2016b). The evaluation is entirely based on a Java software library called HESML which has been developed by the authors in order to replicate all methods evaluated herein. The new IC models obtain rivaling results as regard the state-of-the-art methods and improve our previous models, whilst the experimental survey allows a detailed and conclusive image of the state of the problem to be drawn by setting the new state of the art and quantifying the main achievements of the last three decades.

Keywords: Intrinsic Information Content models, ontology-based semantic similarity measures, IC-based similarity measures, word similarity benchmark, semantic similarity, concept similarity model, experimental survey.

1 Introduction

The human similarity judgments between concepts underlie most of cognitive capabilities, such as categorization, memory, decision-making, and reasoning, as well as the use and discovery of anologies among others. For this reason, this problem has a lot of applications in Artificial Intelligence (AI) and many other related fields. The main research problem studied herein is the proposal of new Information Content (IC) models for ontology-based semantic similarity measures with the aim of estimating the degree of similarity between words as perceived by a human being. However, because of that the common approach to compute word similarity measures is to select the highest pairwise similarity value between the concept sets evoked by each word, our main research problem is closely related to the proposal of concept similarity models, whose aim is to estimate the degree of similarity between concepts instead of words. A concept similarity model is a function $sim : C \times C \rightarrow \mathbb{R}$ defined on a set of concepts which estimates the degree of similarity between concepts as perceived by a human being. The research into concept similarity models, so called in a broad sense as the human similarity judgment problem in cognitive sciences, has given rise to different strategies to tackle the problem of which the ontology-based simi-

Cite this work as: Lastra-Díaz, J. J., and García-Serrano,1A. (2016). A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet. Technical Report TR-2016-01. NLP and IR Research Group. ETSI Informática. Universidad Nacional de Educación a Distancia (UNED).
http://e-spacio.uned.es/fez/view/bibliuned:DptoLSI-ETSI-Informes-Jlastra-refinement

larity measures have proven to be the most successful of them.

The research into ontology-based semantic similarity measures is an old problem in AI and other related fields, such as cognitive psychology Tversky (1977), Natural Language Processing (NLP) and Information Retrieval (IR), Rada et al. (1989). A plethora of ontology-based similarity measures have been proposed in the literature, giving rise to a large set of applications in the fields of NLP, IR, bioengineering and genomics. For instance, Lastra-Díaz (2014) introduces an ontology-based IR model disclosed by Lastra Díaz and García Serrano (2014) which is based on the weighted Jiang-Conrath (J&C) distance introduced and evaluated in Lastra-Díaz and García-Serrano (2015b). Patwardhan et al. (2003) introduce a Word Sense Disambiguation (WSD) method based on the distributional hypothesis and the use of ontology-based similarity measures in order to select the closest evocated concept between a disambiguated word and its neighboring words. Mihalcea et al. (2006) propose a text similarity measure based on the combination of an Inverse Document Frequency (IDF) weighting scheme with any ontology-based similarity measure, which is evaluated in a Paraphrase Detection (PD) task, whilst Fernando and Stevenson (2008) propose a paraphrase detection method based on a quadratic form between Boolean occurrence vectors whose matrix is defined by any ontology-based similarity measure between words. In document clustering, Song et al. (2009) propose a genetic algorithm for text clustering based on a Li et al. (2003) similarity measure, whilst Dagher and Fung (2013) introduce a document clustering method based on a VSM model and a WordNet-based term expansion based on the Jiang and Conrath (1997) distance. Liu et al. (2009) introduce a method for the discovery of relevant WDSL-specified web services based on a WDSL similarity metric defined by the dot product between the provider and query vectors, whose weights are derived from the Li et al. (2003) similarity measure. Martínez et al. (2010) introduce a document anonymization method based on ontology-based similarity measures. Cross and Hu (2011) introduce a semantic alignment quality measure for the Ontology Alignment (OA) problem which relies on the difference between the similarity measure between the concepts in the base ontology and their image in the target ontology; and Pirró and Talia (2010) introduce an ontology mapping method based on a reformulation of the Jiang and Conrath (J&C) distance and the Seco et al. (2004) IC model, whilst Jeong et al. (2008) propose a framework for XML-schema matching based on ontology-based similarity measures. In Oliva et al. (2011), Lee (2011) and Hadj Taieb et al. (2015), the authors introduce different methods for sentence similarity based on ontology-based similarity measures. Other works use similarity measures for the extraction of domain ontologies from the Internet like Wang and Zhou (2009), or from text corpora like Meijer et al. (2014). Montani et al. (2015) propose an ontology-based process similarity metric for process mining that relies on the Wu and Palmer (1994) similarity measure. In the field of bioengineering, Couto et al. (2007) introduce a reformulation of three classic IC-based similarity measures with the aim of computing similarity measures based on the Gene Ontology (GO), whilst Chaves-González and Martínez-Gil (2013) introduce a similarity-based evolutionary method for synonym recognition in the biomedical domain. Other specific similarity measures have been studied for biomedical text mining, such as Pedersen et al. (2007) and Sánchez and Batet (2011), as well as other genomics applications, such as protein function prediction Pesquita et al. (2009), Couto and Pinto (2013) and pathway prediction Chiang et al. (2008).

1.1 The context of our research

An ontology-based semantic similarity measure is a binary concept-valued function $sim : C \times C \rightarrow \mathbb{R}$ defined on a single-root taxonomy of concepts $(C, \leq_C)$ which returns the degree of similarity between concepts as perceived by a human being. Modern research into the problem starts with the pioneering works by Tversky (1977) and Rada et al. (1989) in the fields of cognitive psychology and IR respectively. Tversky (1977) introduce a feature-based similarity measure which requires a representation of the concepts as feature sets, whilst Rada et al. (1989) introduce a semantic distance defined as the length of the shortest path between concepts in a taxonomy. The main drawback of the Rada et al. (1989) measure, as well as other similarity measures which use the length of the shortest path between concepts, is that all the edges in the taxonomy contribute to the overall distance with the same weight, the so-called *uniform weighting* problem. In order to bridge this latter gap, Resnik (1995) introduces the first similarity measure based on an Information Content (IC) model derived from corpus statistics, as well as the first method to compute an IC model, such as those proposed herein.

Every IC-based similarity measure needs a complementary concept-valued function, called the Information Content (IC) model. Given a taxonomy of concepts defined by a triplet $\mathcal{C} = ((C, \leq_C), \Gamma)$ where $\Gamma \in C$ is the supreme element called the root, an Information Content model is a function $IC : C \rightarrow \mathbb{R}^+ \cup \{0\}$, which represents an estimation of the information content for every concept, defined by $IC(c_i) = -log_2(p(c_i))$, $p(c_i)$ being the occurrence probability of each concept $c_i \in C$. Every IC model must satisfy two further properties: (1) nullity in the root, such that $IC(\Gamma) = 0$, and (2) growing monotonicity from the root to the leaf concepts, such that $\forall c_i \leq_C c_j \Rightarrow IC(c_i) \geq IC(c_j)$. Once the IC-based measure is chosen, the IC model is mainly responsible for the definition of the notion of similarity and distance between concepts. Other works, such as Pirró and Euzenat (2010), have also proposed intrinsic IC models for semantic relatedness measures which rely on the whole set of semantic relationships encoded into an ontology.

The first known IC model is based on corpus statistics, which was introduced by Resnik (1995) and detailed in Resnik (1999). The main drawback of the corpus-based IC models is the difficulty in getting a well-balanced and disambiguated corpus for the estimation of the concept probabilities. To bridge this gap, Seco et al. (2004) intro-

duced the first intrinsic IC model in the literature, whose core hypothesis is that the IC models can be directly computed from intrinsic taxonomical features. Therefore, the development of new intrinsic IC-based similarity measures is divided into two subproblems: (1) the proposal of new intrinsic IC models, as in our work, and (2) the proposal of new IC-based similarity measures. In another recent work Lastra-Díaz and García-Serrano (2015a), we introduce a new family of intrinsic and corpus-based IC models called *well-founded IC models*, which is based on the proposal of different methods for the estimation of the conditional probabilities between child and parent concepts within a taxonomy. The main idea behind the new family of *well-founded IC models* is that any IC model should satisfy a set of axioms that algebraically link the conditional probabilities, probability function and IC model in order to define a well-founded probability space.

1.2 Motivation and hypotheses

The first motivation is the finding of two drawbacks in the algorithm to compute the family of *well-founded IC models* introduced in Lastra-Díaz and García-Serrano (2015a). First, the two intrinsic and cognitive IC models called *CondProbLogistic* and *CondProbCosine* do not satisfy the axiom that constrains the sum of probabilities on the leaf nodes to be 1. It is a consequence of the non-linear transformations applied to the conditional probabilities of these two models, a fact that was already mentioned in our aforementioned work. Second, in some cases, the ontologies with multiple inheritance could prevent the IC model satisfying the *growing monotonicity axiom* in concepts with multiple parents. This latest fact means that for some concept pairs $c_i, c_j \in C$, the constraint $c_i \leq_C c_j \Rightarrow IC(c_i) \geq IC(c_j)$ could be violated. In appendix B of our aforementioned work, we prove that the recovery algorithm based on the recursive formula in equation (3) is a sufficient condition for the sum of probabilities over the leaf nodes to be 1, what follows the underlying probability space is well-defined. However, if the taxonomy exhibits multiple inheritance, the probabilities $p(c_i)$ derived from equation (3) could be higher than the probability of any direct parent in some nodes with multiple parents, thus, leading to a violation of the aforementioned growing monotonicity axiom. Our main hypothesis is that the solution to these two drawbacks could lead us to an improvement in the performance of the family of well-founded IC models, in addition to fixing an algebraic inconsistency that moves the family of *well-founded IC model* away from their original design principles.

Second motivation of this work is the lack of an updated and exhaustive evaluation of ontology-based similarity measures and IC models in WordNet, as well as the lack of an exhaustive pairwise statistical significance analysis between them. In the literature, we find some out-of-date similarity benchmarks such as that reported by Budanitsky and Hirst (2001) and Budanitsky and Hirst (2006), and others, more recent but not exhaustive, such as Hadj Taieb et al. (2014b). The largest and most recent word similarity benchmarks in WordNet are introduced by Lastra-Díaz and García-Serrano (2015a) and Lastra-Díaz and García-Serrano (2015b). However, not all of the hybrid IC-based similarity measures evaluated in the latest work have been previously evaluated with many IC models considered herein and the datasets introduced by Miller and Charles (1991), Agirre et al. (2009) and Hill et al. (2015). In addition, most ontology-based similarity measures have never been compared through a statistical significance analysis. Therefore, in the light of the results reported by Lastra-Díaz and García-Serrano (2015a), and in order to provide a conclusive image of the current state of the problem, we introduce herein a new and larger evaluation of IC models and ontology-based similarity measures than those available in the literature. This new evaluation is based on the most recently available datasets and our own software implementation of all the IC models and similarity measures evaluated herein, covering most developments from the pioneering works of Rada et al. (1989) and Seco et al. (2004).

Finally, the last motivation is the replication of previous methods and experiments. Most works introducing similarity measures or IC models during the last decade have only implemented or evaluated classic IC-based similarity measures, such as the Resnik, Lin and Jiang-Conrath measures, avoiding the replication of IC models and similarity measures introduced by other researchers. Some works have not included all the details of their methods, or the experimental setup to obtain the published results, thus, preventing their reproducibility. Most works have copied results published by others. This latest fact has prevented the valuable confirmation of previous methods and results reported in the literature, which is an essential feature of science. Pedersen (2008a), and subsequently Fokkens et al. (2013), warn of the need to reproduce and validate previous methods and results reported in the literature, a suggestion that we subscribe to in our aforementioned works, where we also warn of finding some contradictory results. This replication problem is especially significant in the current state of the problem, in which there is no convincing winner within the family of intrinsic IC-based similarity measures and the performance margin is very narrow, as concluded in our aforementioned works. In addition, Pedersen (2008a) also warns of the need of releasing the software developed for the evaluation of new methods and experiments reported in the literature with the aim of allowing their reproducibility. Following the suggestions from Pedersen, we introduce our new software library of ontology-based semantic similarity measures and IC models together with a set of reproducible experiments in a forthcoming paper, Lastra-Díaz and García-Serrano (2016).

The proposed refinements close the algebraic and algorithmic definition of the family of *well-founded IC models*, giving rise to research into further IC models within this family.

For the experimental survey, our main hypotheses are as follows:

H1. A group of recent IC-based similarity measures outperform the path-based similarity measures, as well

as the classic IC-based measures, but there is no statistically significant difference between them.

H2. There is no statistically significant difference in performance between most intrinsic IC models and the best performing corpus-based IC model defined as baseline, which is derived from the *"ic-treebank-add1.dat"* file in the Pedersen (2008b) dataset.

H3. A small set of the best performing intrinsic IC models outperform the best performing corpus-based IC model defined as baseline.

H4. The classic IC-based similarity measures proposed by Resnik, Jiang and Conrath, and Lin have been definitively outperformed by a small set of state-of-the-art IC-based similarity measures.

H5. The practical use of the current hybrid IC-based similarity measures that are based on the length of the shortest path is prevented by their high computational cost in comparison with the other methods with a similar performance.

H6. Most IC-based similarity measures perform better with a specific IC model.

H7. The state-of-the-art IC-based similarity measures outperform the best corpus-based similarity measures in the SimLex665 dataset.

H8. The proposed refinement into the computation method of the well-founded IC models could lead us to an improvement in their performance.

1.3 Research problem and contributions

The main aims of this paper are as follows. First, the proposal of a refinement into the four-step algorithm used to compute the family of well-founded IC models with the aim of eliminating the aforementioned drawbacks of the computational method introduced in our previous work, Lastra-Díaz and García-Serrano (2015a). Second, the proposal of eight new intrinsic IC models and one new corpus-based IC model in the new framework of our family of well-founded IC models. And third, the introduction of a new and very detailed experimental survey of IC models and ontology-based similarity measures on WordNet with a complete detailed statistical significance analysis between IC models and similarity measures, including the evaluation of most ontology-based similarity measures since the work of Rada et al. (1989) and all WordNet-based IC models reported in the literature, with the only exception of the IC models recently introduced by Harispe et al. (2015a) and Ben Aouicha et al. (2016b).

The refinement of the *well-founded IC models* allows a new family of IC models to be derived from the previous models introduced by Lastra-Díaz and García-Serrano (2015a), as well as three new strategies to compute the conditional probabilities. The new intrinsic IC models are called *CondProbRefHyponyms, CondProbRefUniform, CondProbRefLeaves, CondProbRefLogistic*, *CondProbRefCosine*, *CondProbRefLogisticLeaves*, *CondProbRefCosineLeaves* and *CondProbRefLeavesSubsumersRatio*, whilst the new corpus-based IC model is called *CondProbRefCorpus.* The *CondProbRefLeavesSubsumersRatio* IC model is a reformulation of the Sánchez et al. (2011) IC model in the framework defined by our family of IC models.

The new experimental survey includes most of the intrinsic and corpus-based IC models evaluated in Lastra-Díaz and García-Serrano (2015a), as well as the nine new IC models introduced herein, one of the unexplored intrinsic IC models introduced by Blanchard et al. (2008), and most ontology-based similarity measures since the work by Rada et al. (1989). The word similarity benchmarks introduced herein include the five most significant datasets on the problem, as well as a very detailed pairwise statistical significance analysis between the IC models and ontology-based similarity measures. The benchmarks reported herein are, to the best of our knowledge, the largest experimental survey on intrinsic IC models and ontology-based similarity measures on WordNet reported in the literature, which is based on a same code implementation. We exactly reproduce the same experiments from Lastra-Díaz and García-Serrano (2015a), but with a much larger set of IC models and ontology-based similarity measures. Our experiments include a set of the hybrid IC-based similarity measures based on the length of the shortest path between concepts which were evaluated in Lastra-Díaz and García-Serrano (2015b) and subsequently discarded because of their high computational cost. The experimental survey includes 22 ontology-based similarity measures, 22 intrinsic IC models, and 3 corpus-based IC models.

The rest of the paper is structured as follows. Section 2 reviews the literature on concept similarity models. Section 3 summarizes the factual state of the art of the problem, whilst section 3.1 reviews the literature on intrinsic IC models. Section 4 introduces the proposed refinement in the well-founded IC models, as well as the new IC models derived from it. Section 5 describes the evaluation methodology and the results obtained. Section 6 introduces an in-depth discussion of the results. Last section presents our conclusions and future work. Finally, appendix groups the summary data tables and all raw data tables resulting from the evaluation.

2 Concept similarity models

This section makes a comparison between the concept and word similarity models proposed in the literature which we categorize as ontology-based and corpus-based similarity measures, and the most recent concept similarity models proposed in cognitive psychology. First, we compare the main strategies adopted to tackle the problem, and finally, we review the literature on corpus-based and ontology-based similarity measures.

2.1 Comparison of strategies

In the fields of NLP and IR, we find two different types of similarity models to estimate the degree of similarity between words: (1) ontology-based similarity measures as

Reference	Definition of the non IC-based similarity measures
Rada et al. (1989)	$sim_{Rada}(c_1,c_2) = 1 - \frac{1}{2} d_{Rada}(c_1,c_2)$
	$d_{Rada}(c_1,c_2) = len(c_1,c_2) = \min_{\forall \alpha \in Paths_{(c_1,c_2)}} \left\{ \sum_{e_{ij} \in \alpha} 1 \right\}$
Wu and Palmer (1994)	$sim_{W\&P}(c_1,c_2) = \frac{2 \times depth(LCA(c_1,c_2))}{len(c_1,LCA(c_1,c_2)) + len(c_2,LCA(c_1,c_2)) + 2 \times depth(LCA(c_1,c_2))}$
Leacock and Chodorow (1998)	$sim_{L\&C}(c_1,c_2) = -log\left(\frac{1+len(c_1,c_2)}{2 \times maxdepth}\right)$
Li et al. (2003)	$sim_{Li\_s3}(c_1,c_2) = e^{-\alpha * len(c_1,c_2)}, \quad \alpha^* = 0.25$
Li et al. (2003)	$sim_{Li\_s4}(c_1,c_2) = e^{-\alpha * len(c_1,c_2)} \times \frac{e^{\beta * d} - e^{-\beta * d}}{e^{\beta * d} + e^{-\beta * d}}$, $\alpha^* = 0.2$ $\beta^* = 0.6$
	$d = depth(LCA(c_1,c_2))$
Al-Mubaid and Nguyen (2009)	$d_{Mubaid}(c_1,c_2) = log(1 + len(c_1,c_2) * (depthmax - depth(LCS(c_1,c_2))))$
Pedersen et al. (2007)	$sim_{Path}(c_1,c_2) = \frac{1}{1+len(c_1,c_2)}$
Sánchez et al. (2012)	$dis_{S\&B}(c_1,c_2) = log_2\left(1 + \frac{\lvert\phi(c_1) \setminus \phi(c_2)\rvert + \lvert\phi(c2) \setminus \phi(c_1)\rvert}{\lvert\phi(c_1) \setminus \phi(c_2)\rvert + \lvert\phi(c_2) \setminus \phi(c_1)\rvert + \lvert\phi(c_1) \cap \phi(c_2)\rvert}\right)$
	$\phi(a) = \{c \in C \mid a \leq c\}$
Hadj Taieb et al. (2014b)	$sim_{Taieb\_1}(c_1,c_2) = \lvert TermDepth(c_1,c_2)\rvert \times TermHypo(c_1,c_2)$
	$TermDepth(c_1,c_2) = \frac{2 \times depth(c_1,c_2)}{depth(c_1) + depth(c_2)}$
	$TermHypo(c_1,c_2) = \frac{2 \times Spec_{Hypo}(c_1,c_2)}{Spec_{Hypo}(c_1,c_2) + Spec_{Hypo}(c_1,c_2)}$
	$Spec_{Hypo}(c_1,c_2) = 1 - \frac{log(HypoValue(c))}{log(HypoValue(root))}$
	$HypoValue(c) = \sum_{c' \in HypoInc(c)} P(depth(c'))$
	$P(depth(c')) = \frac{\lvert\{c' \in C \mid depth(c') = depth(c)\}\rvert}{\lvert C \rvert}$
	$depth(c) =$ length of the longest ascending path $c \longrightarrow root$
	$HypoInc(c) = \{c' \in C \mid c' \leq c\}$

Table 1: State-of-the-art non IC-based similarity measures evaluated in our experiments.

in our work, and (2) corpus-based similarity and relatedness measures. The ontology-based similarity measures are based on the definition of binary concept-valued similarity functions on "is-a" taxonomies, which have proven in Lastra-Díaz and García-Serrano (2015a) to be the best approximation to similarity human judgments on the noun subset of the SimLex dataset Hill et al. (2015), as being efficient, robust and easy to implement. However, the main drawback of the ontology-based similarity measures is the limited coverage of the ontologies and the cost and difficulties of building them. Other drawback of the ontology-based methods is the requirement of a single taxonomy that includes all the words to be compared, although this problem has given rise to the proposal of methods for the estimation of semantic similarity measures combining multiple ontologies, such as the general-purpose method introduced by Al-Mubaid and Nguyen (2009), the method for feature-based measures proposed by Solé-Ribalta et al. (2014) and the method for IC-based similarity measures proposed by Batet et al. (2014). On the other hand, the corpus-based similarity and relatedness measures mainly rely on the distributional hypothesis, and they are commonly based on the statistical co-occurrence between word contexts in large corpora, as a means of estimating the degree of similarity between words. The corpus-based measures "can confuse similarity with relatedness" (Li et al., 2015, §1). In addition, "it is commonly considered that distributional measures can only be used to capture semantic relatedness" (Harispe et al., 2015b, §2.5.2), and "they have traditionally performed poorly when compared to WordNet-based measures" (Mohammad and Hirst, 2012, p.1). This latter fact is confirmed by the recent comparisons between ontology-based and corpus-based similarity measures reported by (Banjade et al., 2015, Table 1) and Le and Fokkens (2015), as well as our benchmarks in (Lastra-Díaz and García-Serrano, 2015a, §6.4). It is worth to note that the ontology-based similarity measures use an explicitly defined concept similarity model with the aim of estimating the degree of similarity between words whose specific meaning (evocated concept) is unknown, whilst the corpus-based measures use the occurrence of the words in a specific context, whose meaning (concept) is implicitly defined by the context.

Finally, the research into the similarity judgments problem in cognitive psychology derives from the pioneering work of Tversky (1977). The research into the field of IR has focused on the proposal of a plethora of symmetric and contextless similarity measures guided by experimental evaluation. On the contrary, the research into cognitive sciences has followed a parallel line more focused on the definition of theoretical models capable of explaining several non-metric phenomena in the human similarity judgments described by Tversky (1977) and Pothos et al. (2015), such as: (1) asymmetry or non-commutativity, (2) context dependency and (3) the conjunction fallacy. The most recent cognitive similarity model is introduced by Pothos et al. (2013) and Pothos and Trueblood (2015), being inspired by a quantum probability approach for cognition proposed by Busemeyer and Bruza (2012), whose non-commutative nature allows the representation of different non-metric phenomena. However, the quantum probability similarity model has not yet been experimentally evaluated.

2.2 Corpus-based measures

Many corpus-based similarity or relatedness measures are based on concept-based resources, such as Wikipedia. For instance, Strube and Ponzetto (2006) introduce WikiRelate, a method for computing the semantic relatedness between words based on a graph derived from Wikipedia. WikiRelate extracts the Wikipedia pages associated to each input word and builds a taxonomy of categories by merging the categories that the pages belong to. Finally, WikiRelate uses standard path-based and IC-based similarity measures on the recovered taxonomy in order to compute the relatedness measure between words. We can interpret WikiRelate as a two-stage method based on the combination of a taxonomy recovering method, such as the method recently proposed by Ben Aouicha et al. (2016a), with any standard ontology-based similarity measure. Gabrilovich and Markovitch (2007) introduce a semantic relatedness method for word and documents, called ESA, which represents the meaning of a word or text as a weighted vector of Wikipedia concepts (articles); whilst Agirre et al. (2009) introduce several distributional relatedness measures based on a vector space model trained on a large Web corpus, which favourably compare with a large set of ontology-based similarity measures on WordNet.

On the other hand, another very active line of research in corpus-based similarity measures is the proposal for hybrid concept-based distributional measures, which integrate knowledge bases (KBs) or explicit "is-a" semantic networks in order to overcome the lack of well-defined semantic knowledge. For instance, Patwardhan and Pedersen (2006) introduce a similarity and relatedness measure which relies on the gloss vector overlapping between the extended WordNet gloss vectors of two input concepts. Mohammad and Hirst (2006) introduce a hybrid distributional measure which relies on the cosine function and the concept-based conditional probabilities for the words derived from the Roget's thesaurus. Alvarez and Lim (2007) propose a hybrid distributional similarity measure that relies on the product of two taxonomical WordNet-based functions with a gloss overlapping factor by using "is-a" and "part-of" relationships, whilst Li et al. (2015) introduce another hybrid distributional measure whose core idea is that the similarity computation relies on truly "is-a" relationships, which are derived from a very large web corpus by using an automatic method based on syntactic rules.

Other family of relatedness measures are based on randow walks on weighted graphs derived from different knowledges sources, such as Wikipedia and WordNet. For instance, Hughes and Ramage (2007) propose a semantic relatedness measure between word pairs which is based on a random walk using Personalized PageRank on a weighted graph derived from WordNet and corpus statistics, whilst Yeh et al. (2009) extend their previous work on semantic relatedness measures based on random walks to Wikipedia, and Ramage et al. (2009) propose a corpus-based measure based on a random walk on WordNet with the aim of estimating the semantic similarity between text fragments. Finally, Yazdani and Popescu-Belis (2013) propose a method for estimating the semantic relatedness between concepts based on a random walk approach on a Wikipedia concept network with two link types: the hypertext links between Wikipedia articles (concepts), and the lexical similarity between them defined by the cosine score between the vectors representing each article.

Another growing research trend on corpus-based semantic similarity and relatedness measures is the development of word embeddings, such as those proposed by Mikolov et al. (2013), Pennington et al. (2014) and Suzuki and Nagata (2015), whose core idea is the learning of a vector representation (embedding) for large vocabularies, such that the Euclidean distance between word vectors reflects their semantic similarity. Most word embeddings use a large corpora in their learning process, thus, they are a subfamily of the corpus-based methods. The word embedding methods commonly use complex machine learning algorithms, which are time-consuming and hard to reproduce. However, once the vector representations are computed, their evaluation mainly depends on the dimensionality of the vector space, thus, they can be very efficient for large vocabularies and low dimensionality.

2.3 Ontology-based similarity measures

In two recent works, Lastra-Díaz and García-Serrano (2015b) and Lastra-Díaz and García-Serrano (2015a), we provide a very detailed review of the current ontology-based semantic measures, thus, we only provide herein a categorization in order to introduce the similarity measures that will be evaluated in our experiments. For a more in-depth review of the topic, we refer the reader to our aforementioned works, especially the former, and the recent book by Harispe et al. (2015b).

We categorize the current ontology-based semantic measures into four subfamilies as follows: (1) edge-counting similarity measures, the so called path-based measures, whose core idea is the use of the length of the shortest path between concepts as an estimation of their degree of similarity, such as the pioneering work of Rada et al. (1989) and the subsequent works of Wu and Palmer (1994), Leacock and Chodorow (1998), Hirst and St-Onge (1998), Pedersen et al. (2007) and Al-Mubaid and Nguyen (2009); (2) IC-based similarity measures whose core idea is the use of an Information Content (IC) model, such as the pioneering work of Resnik (1995), and the measures proposed by Jiang and Conrath (1997) and Lin (1998); (3) feature-based measures, whose core idea is the use of set-theory operators between the feature sets of the concepts, such as the pioneering work of Tversky (1977), and more recently Sánchez et al. (2012), whose core idea is the use of the overlapping of ancestor sets as an estimation of the overlapping between the unknown feature sets of the concepts; and finally, (4) other similarity measures that cannot be directly categorized into any previous family, which are based on taxonomical features derived from set-theory operators Batet et al. (2011), or novel contributions of the hyponym set Hadj Taieb et al. (2014b). Out of our previous categorization, it was also worth mentioning some proposals of *aggregated similarity measures*, such as Martinez-Gil (2016), whose key

Classic IC-based similarity measures	
Resnik (1995)	$sim_{Resnik}(c_1, c_2) = IC(MICA(c_1, c_2))$
Jiang and Conrath (1997)	$d_{J\&C}(c_1, c_2) = IC(c_1) + IC(c_2) - 2IC(MICA(c_1, c_2))$ $sim_{J\&C}(c_1, c_2) = 1 - \frac{1}{2} d_{J\&C}(c_1, c_2)$
Lin (1998)	$sim_{Lin}(c_1, c_2) = \frac{2IC(MICA(c_1, c_2))}{IC(c_1)+IC(c_2)}$
IC-based reformulations of the Tversky similarity measure	
Pirró and Seco (2008)	$sim_{P\&S}(c_1, c_2) = \begin{cases} 3IC(MICA(c_1, c_2)) - IC(c_1) - IC(c_2) & \text{, if } c_1 \neq c_2 \\ 1 & \text{, if } c_1 = c_2 \end{cases}$
Monotone transformations of classic IC-based similarity measures	
Pirró and Euzenat (2010)	$sim_{FaITH}(c_1, c_2) = \frac{IC(MICA(c_1, c_2))}{IC(c_1)+IC(c_2)-IC(MICA(c_1, c_2))}$
Meng and Gu (2012)	$sim_{Meng}(c_1, c_2) = e^{sim_{Lin}(c_1, c_2)} - 1 = e^{\frac{2IC(MICA(c_1, c_2))}{IC(c_1)+IC(c_2)}} - 1$
Garla and Brandt (2012)	$sim_{path\_IC}(c_1, c_2) = \frac{1}{1+d_{J\&C}(c_1, c_2)}$
Lastra-Díaz and García-Serrano (2015b)	$sim_{cosJ\&C}(c_1, c_2) = 1 - cos\left(\frac{\pi}{2}\left(1 - \frac{d_{J\&C}(c_1, c_2)}{2 * max_{d_{J\&C}}}\right)\right)$ $max_{d_{J\&C}} = \max_{c \in Leaves(C)} \{IC(c)\}$
Hybrid IC-based similarity measures based on the shortest path length	
Li et al. (2003)	$sim_{Li\_s9}(c_1, c_2) = sim_{Li\_s4}(c_1, c_2) * \frac{e^{\lambda * IC} - e^{-\lambda * IC}}{e^{\lambda * IC} + e^{-\lambda * IC}}$, $\lambda^* = 0.4$ $IC = MICA(c_1, c_2)$
Zhou et al. (2008b)	$sim_{Zh}(c_1, c_2) = 1 - k \times \left(\frac{log(len(c_1, c_2)+1)}{log\left(2 * \max_{c \in T}\{depth(c)\} - 1\right)}\right)$ $- \frac{1}{2}(1-k) \times d_{J\&C}(c_1, c_2)$ $\quad k^* = \frac{1}{2}$ by default
Meng et al. (2014)	$sim_{Meng2014}(c_1, c_2) = sim_{Lin}(c_1, c_2)^{\left(\frac{1-e^{-k*len(c_1, c_2)}}{e^{-k*len(c_1, c_2)}}\right)}$, $k^* = 0.08$ $sim_{Gao}(c_1, c_2) = e^{-\alpha L(c_1, c_2)}$, $\alpha^* = 0.15$ and $\beta^* = 2.05$ $L(c_1, c_2) = wt(c_1, c_2) * len(c_1, c_2)$
Gao et al. (2015)	$wt = \begin{cases} \left(\frac{1+IC(MICA(c_1, c_2))}{IC(MICA(c_1, c_2))}\right)^{\beta} & , IC(MICA(c_1, c_2)) \geq 1 \\ 2^{\beta} & , 1 > IC(MICA(c_1, c_2)) \geq 0 \end{cases}$ $sim_{coswJ\&C}(c_1, c_2) = 1 - cos\left(\frac{\pi}{2}\left(1 - \frac{d_{wJ\&C}(c_1, c_2)}{2 * max_{d_{J\&C}}}\right)\right)$
Lastra-Díaz and García-Serrano (2015b)	$d_{wJ\&C}(c_1, c_2) = \min_{\forall \alpha \in Paths_{(c_1, c_2)}} \left\{\sum_{e_{ij} \in \alpha} w(e_{ij})\right\}$ $w(e_{ij}) = \begin{cases} -log_2(p(c_i \vert c_j)) & \text{, if } p(c_i \vert c_j) \text{ are known} \\ \vert IC(c_i) - IC(c_j)\vert & \text{, otherwise} \end{cases}$

Table 2: Definition of the state-of-the-art IC-based similarity measures evaluated in our experiments.

feature is the merging of multiple ontology-based similarity measures in order to produce a final similarity judgement.

In addition to the four subfamilies of ontology-based similarity measures aforementioned above, we categorize the family of IC-based similarity measures into the following four subgroups, as shown in table 2: (1) the first group of classic IC-based measures made up of the similarity measures introduced by Resnik (1995), Jiang and Conrath (1997) and Lin (1998); (2) a second group that we call hybrid or path-based IC-based similarity measures, which is defined by those measures that make up an IC model with any function based on the length of the shortest path between concepts, such as the pioneering work of Li et al. (2003), and other subsequent works such as Zhou et al. (2008a), Meng et al. (2014), Gao et al. (2015), and the two weighted IC-based similarity measures introduced by Lastra-Díaz and García-Serrano (2015b); (3) a third group that is based on any reformulation strategy between different approaches, such as the IC-based reformulations of the Tversky measure in Pirró (2009) and Pirró and Euzenat (2010), as well as the IC-based reformulation of most edge-counting methods introduced by Sánchez and Batet (2011); and finally, (4) a fourth group that is based on a monotone transformation of any classic IC-based similarity measure, such as the exponential-like scaling of the Lin (1998) measure introduced by Meng and Gu (2012), the reciprocal of the J&C distance introduced by Garla and Brandt (2012), and another cosine-based normalization of the J&C distance introduced by Lastra-Díaz and García-Serrano (2015b). In addition, we show herein that the FaITH similarity measure introduced by Pirró and Euzenat (2010) is also a monotone transformation of the Lin (1998) similarity measure, despite its initial design being based on a reformulation of the Tversky (1977) measure. Table 3 shows

Rada et al. (1989) similarity measure and its monotone transformations

Rada et al. (1989)	$sim_{Rada}(c_1,c_2) = 1 - \frac{1}{2}d_{Rada}(c_1,c_2)$
	$d_{Rada}(c_1,c_2) = len(c_1,c_2) = \min_{\forall\alpha\in Paths_{(c_1,c_2)}} \left\{ \sum_{e_{ij}\in\alpha} 1 \right\}$
	$sim_{L\&C}(c_1,c_2) = -log\left(\frac{1+len(c_1,c_2)}{2\times maxdepth}\right)$
Leacock and Chodorow (1998)	Factorization:
	$sim_{L\&C}(c_1,c_2) = \varphi(x) \circ sim_{Rada}(c_1,c_2)$
	$\varphi(x) = -log\left(\frac{3-2x}{2\times maxdepth}\right)$
Li et al. (2003)	$sim_{Li\_s3}(c_1,c_2) = e^{-\alpha * len(c_1,c_2)}, \quad \alpha^* = 0.25$
	Factorization:
	$sim_{Li\_s3}(c_1,c_2) = \varphi(x) \circ sim_{Rada}(c_1,c_2)$
	$\varphi(x) = e^{2\alpha*(x-1)} \overset{\alpha^*=0.25}{\longrightarrow} \varphi^*(x) = e^{\frac{(x-1)}{2}}$
Pedersen et al. (2007)	$sim_{Path}(c_1,c_2) = \frac{1}{1+len(c_1,c_2)}$
	Factorization:
	$sim_{Path}(c_1,c_2) = \varphi(x) \circ sim_{Rada}(c_1,c_2)$
	$\varphi(x) = \frac{1}{3-2x}$

Lin (1997) similarity measure and its monotone transformations

Lin (1998)	$sim_{Lin}(c_1,c_2) = \frac{2IC(MICA(c_1,c_2))}{IC(c_1)+IC(c_2)}$
	$sim_{FaITH}(c_1,c_2) = \frac{IC(MICA(c_1,c_2))}{IC(c_1)+IC(c_2)-IC(MICA(c_1,c_2))}$
Pirró and Euzenat (2010)	Factorization:
	$sim_{FaITH}(c_1,c_2) = \varphi(x) \circ sim_{Lin}(c_1,c_2)$
	$\varphi(x) = \frac{x}{2-x}$
Meng and Gu (2012)	$sim_{Meng}(c_1,c_2) = e^{sim_{Lin}(c_1,c_2)} - 1$
	Factorization:
	$sim_{Meng}(c_1,c_2) = \varphi(x) \circ sim_{Lin}(c_1,c_2)$
	$\varphi(x) = e^x - 1$

Jiang and Conrath (1997) similarity measure and its monotone transformations

Jiang and Conrath (1997)	$d_{J\&C}(c_1,c_2) = IC(c_1) + IC(c_2) - 2IC(MICA(c_1,c_2))$
	$sim_{J\&C}(c_1,c_2) = 1 - \frac{1}{2}d_{J\&C}(c_1,c_2)$
	$sim_{path\_IC}(c_1,c_2) = \frac{1}{1+d_{J\&C}(c_1,c_2)}$
Garla and Brandt (2012)	Factorization:
	$sim_{path\_IC}(c_1,c_2) = \varphi(x) \circ sim_{J\&C}(c_1,c_2)$
	$\varphi(x) = 3 - 2x$
	$sim_{cosJ\&C}(c_1,c_2) = 1 - cos\left(\frac{\pi}{2}\left(1 - \frac{d_{J\&C}(c_1,c_2)}{2*max_{d_{J\&C}}}\right)\right)$
	$max_{d_{J\&C}} = \max_{c\in Leaves(C)} \{IC(c)\}$
Lastra-Díaz and García-Serrano (2015b)	Factorization:
	$sim_{cosJ\&C}(c_1,c_2) = \varphi(x) \circ \phi(t) \circ sim_{J\&C}(c_1,c_2)$
	$\varphi(x) = 1 - cos\left(\frac{\pi}{2}x\right)$
	$\phi(t) = 1 - \frac{1-t}{max_{d_{J\&C}}}$, normalization function

Table 3: Equivalence classes of similarity measures induced by any monotone transformation from any classic similarity measure.

the monotonicity relationships between most IC-based similarity measures which have been experimentally confirmed in our evaluation. For the sake of completeness of our experimental survey, we also evaluate herein all non IC-based similarity measures shown in table 1, despite the present work is focused on new IC models and their evaluation with the state-of-the-art IC-based similarity measures shown in table 2.

Finally, Stanchev (2014) introduces a similarity graph from WordNet with the aim of computing the similarity between words. In addition to the taxonomical structure from WordNet, the graph uses the definition and examples of use of the WordNet concepts as evidence on the relationships between concepts. The similarity graph is defined by a collection of oriented edges with asymmetric weights, in which the weights between parent and child concepts encode the probability that a user interested in the source node of an edge is also interested in the concept associated to the destination node. The similarity measure is defined as the product of the edge weights throughout the path between the word nodes. Despite some weights being defined in an arbitrary way, the method obtains outstanding results in the Miller and Charles (1991) dataset, and introduces for the first time an asymmetrical path-based method founded on probability theory. We note that the similarity measure introduced by Stanchev is closely related to our weighted J&C distance, denoted by $dwJ\&C$ in table 2, as our measure matches the logarithm of the product of conditional probabilities between the word nodes. However, the basic form of the $dwJ\&C$ distance does not integrate the word nodes into the WordNet taxonomy and the weights are symmetric, the edge weights being the logarithm of the conditional probabilities.

2.4 Summary and positioning

In summary, the ontology-based similarity measures are efficient, easy to implement and more accurate than the corpus-based methods, whilst the corpus-based measures offer a broader lexical coverage at the expense of a high complexity and computational cost, as well as the difficulties to obtain well-balanced learning corpus. However, the corpus-based relatedness measures based on word embeddings combine the broad coverage of the corpus-based methods with an efficient evaluation method in operation mode. On the other hand, unlike the theoretical models developed in cognitive psychology which have not yet evaluated, the ontology-based similarity measures have been successfully evaluated in many human similarity benchmarks, and they have contributed to the development of a large set of applications. For these reasons, we are focusing our research effort on the development of new IC models and ontology-based similarity measures.

3 State of the art

This section summarizes the current factual state of the art on ontology-based similarity measures and IC models and review the related work on IC models.

The state of the art in ontology-based similarity measures is defined by the family of intrinsic IC-based measures, which are defined by the combination of one specific IC-based similarity measure with any intrinsic IC model. More specifically, our cosine-normalized Jiang-Conrath ($cosJ\&C$) similarity measure is currently the best performing ontology-based similarity measure according to the evaluation on the five most significant datasets reported in (Lastra-Díaz and García-Serrano, 2015a, table 6). However, in this latest work we did not evaluate other hybrid IC-based measures that obtained state-of-the-art results in Lastra-Díaz and García-Serrano (2015b), such as our hybrid measure $coswJ\&C$ and the Zhou et al. (2008b) similarity measure. Likewise, the $cosJ\&C$ similarity measure is the only measure that obtains a statistically significant higher performance than the baseline, (Lastra-Díaz and García-Serrano, 2015a, fig.3). However, we also prove that there is no statistically significant difference between the $cosJ\&C$ similarity measure and those introduced by Meng and Gu (2012) and Pirró and Euzenat (2010).

The outperformance of the IC-based similarity measures is supported by several recent WordNet-based benchmarks, such as Lastra-Díaz and García-Serrano (2015a), Lastra-Díaz and García-Serrano (2015b) and Hadj Taieb et al. (2014b), as well as other older ones, such as Budanitsky and Hirst (2006), Pirró (2009) and Sánchez et al. (2011). Another benchmark in bioengineering introduced by Garla and Brandt (2012) also confirms the outperformance of an intrinsic IC-based similarity measure derived from the reciprocal of the J&C distance. Likewise, McInnes and Pedersen (2013) prove the outperformance of the classic IC-based similarity measures over the path-based measures and gloss-based relatedness measures in a WSD benchmark in bioengineering, but it is also proven that there is no a statistically significant difference between a corpus-based IC model and the intrinsic IC model introduced by Sánchez et al. (2011). This latest conclusion on the debate between intrinsic and corpus-based IC models is endorsed in a more conclusive manner by the recent benchmarks in our aforementioned works.

In our aforementioned works, we conclusively prove several significant facts on the state of the art of IC models as follows. First, contrary to what the research community thought, most corpus-based IC models derived from the unexplored "*.add1" set of WordNet-based frequency files in Pedersen (2008b) rival the state-of-the-art intrinsic IC models, (Lastra-Díaz and García-Serrano, 2015b, table 6). Second, the best performing IC model on average is the Seco et al. (2004) IC model, (Lastra-Díaz and García-Serrano, 2015a, table 5). Third, there is no a statistical significant difference between most state-of-the-art intrinsic IC models, as well as between most intrinsic IC models and the baseline IC model defined by a corpus-based IC model derived from the "ic-treebank-add1.dat" file in the aforementioned Pedersen dataset, (Lastra-Díaz and García-Serrano, 2015a, fig.2). And finally, the Sánchez and Batet (2012) IC model is the only one that obtains a statistically significant higher performance than the corpus-based IC model defined as

IC models	Definition
Resnik (1999)	$IC_{Resnik} = -log_2\left(\widehat{p}\left(c_i\right)\right)\ ,\ \widehat{p}\left(c_i\right) = \frac{f(c_i)}{N} = \frac{f(c_i)}{f(\Gamma)}$ $f\left(c_i\right) = TF\left(c_i\right) + IF\left(c_i\right) = TF\left(c_i\right) + \sum_{\forall c_j \mid c_i \in LA(c_j)} f\left(c_j\right)$
Seco et al. (2004)	$IC_{Seco}\left(c\right) = 1 - \frac{log(\lvert Hypo(c)\rvert + 1)}{log(max\_nodes)}$
Zhou et al. (2008a)	$IC_{Zhou}\left(c\right) = k\left(1 - \frac{log(\lvert Hypo(c)\rvert + 1)}{log(max\_nodes)}\right) + (1-k)\frac{log(depth(c))}{log(depth_{max})},\ \ k^* = \frac{1}{2}$
Blanchard et al. (2008)	$IC_g\left(c_i\right) = -log_2\left(\frac{\lvert SubsumedLeaves(c_i)\rvert}{maxLeaves}\right)$ $SubsumedLeaves\left(c_i\right) = \{c_j \in C \mid c_j \leq_C c_i \ \wedge\ c_j \text{ is leaf }\}$
Sánchez et al. (2011)	$IC_{Sánchez2011}\left(c_i\right) = -log_2\left(\frac{\frac{\lvert Leaves(c_i)\rvert}{\lvert subsumers(c_i)\rvert} + 1}{maxLeaves + 1}\right)$ $\begin{cases} Leaves\left(c_i\right) = \{c_j \in C \mid (c_j \leq_C c_i \ \wedge\ c_j \neq c_i) \ \wedge\ c_j \text{ is leaf}\} \\ subsumers\left(c_i\right) = \{c_j \in C \mid c_i \leq_C c_j\} \end{cases}$
Sánchez and Batet (2012)	$IC_{Sánchez2012}\left(c\right) = -log_2\left(\frac{commonness(c)}{commonness(root)}\right)$ $\begin{cases} commonness\left(c\right) = \frac{1}{\lvert Subsmers(c)\rvert} & , c \text{ leaf} \\ commonness\left(c\right) = \sum_{\forall l \mid l \text{ is leaf and } l<c} commonness\left(l\right) & , c \text{ not leaf} \end{cases}$
Meng et al. (2012)	$IC_{Meng}\left(c\right) = \frac{log(depth(c))}{log(depth_{max})} \times \left(1 - \frac{log\left(1 + \sum_{a \in Hypo(c)} \frac{1}{depth(a)}\right)}{log(Node_{max})}\right)$
Yuan et al. (2013)	$IC_{Yuan}\left(c\right) = f_{depth}\left(c\right)\left(1 - f_{leaves}\left(c\right)\right) + f_{hyper}\left(c\right)$ $\begin{cases} f_{depth}\left(c\right) = \frac{log(depth(c))}{log(depth_{max})} \\ f_{leaves}\left(c\right) = \frac{log(\lvert Leaves(c)\rvert + 1)}{log(Leaves_{max} + 1)} \\ f_{hyper}\left(c\right) = \frac{log(\lvert Hyper(c)\rvert + 1)}{log(Node_{max})} \end{cases}$
Hadj Taieb et al. (2014a)	$IC_{Taieb}\left(c\right) = \left(\sum_{a \in HyperInc(c)} Score\left(a\right)\right) \times AvgDepth\left(c\right)$ $AvgDepth\left(c\right) = \frac{1}{\lvert HyperInc(c)\rvert} \times \sum_{c' \in HyperInc(c)} depth\left(c'\right)$ $Score\left(c\right) = \left(\sum_{c' \in DirectHyper(c)} \frac{depth(c')}{\lvert HypoInc(c')\rvert}\right) \times \lvert HypoInc\left(c\right)\rvert$ $HypoInc\left(c\right) = \{a \in C \mid a \leq c\} \quad HyperInc\left(c\right) = \{a \in C \mid c \leq a\}$
Adhikari et al. (2015)	$IC_{Adhikari}\left(c\right) = \frac{log(depth(c)+1)}{log(depthmax+1)} \times \left(1 - log\left(\frac{\frac{\lvert Leaves(c) \times nmih(c)\rvert}{Leaves_{max}}}{\lvert subsmers'(c)\rvert} + 1\right)\right)$ $\times \left(1 - \frac{log\left(1 + \sum_{a \in Hypo(c)} \frac{1}{depth(a)}\right)}{log(Node_{max})}\right) \quad subsmers'\left(c\right) = subsmers\left(c\right) \cup \{c\}$

Table 4: State-of-the-art Information Content models evaluated in our experiments.

baseline, (Lastra-Díaz and García-Serrano, 2015a, fig.2).

In order to overcome the lexical coverage limitation associated to the ontologies, we argue that at least two strategies could be explored. The first strategy is the ontology population based on WordNet by using any automatic WordNet-based semantic annotation method, such as that explored by Sanfilippo et al. (2005). A second strategy is the automatic assembly of broad coverage "is-a" taxonomies from a large corpus such as Wikipedia, as is recently proposed and evaluated by Ben Aouicha et al. (2016a).

Finally, despite the plethora of ontology-based similarity measures and IC models available in the literature, the selection of a specific similarity measure for a particular application is still an open problem. For instance, a recent benchmark in a biomedical ontology-based IR task by Alonso and Contreras (2016) proves that there is no a statistically significant difference in performance between the *intrinsic IC* measure in (Garla and Brandt, 2012, eq. (13)) and the similarity measure introduced by Pedersen et al. (2007). This latter fact questions the extrapolation of the results and conclusions obtained in classic word similarity benchmarks to specific similarity-based applications. Thus, in order to improve our understanding of the problem, we suggest that the evaluation methodology of ontology-based similarity measures should be reconsidered by defining new task-oriented benchmarks and larger datasets. In this latter line of research, Jurgens et al. (2015) introduce a new similarity evaluation method called Cross-Level Semantic Similarity (CLSS), whose aim is to measure the contribution of the degree of similarity between small language units to the semantic similarity between larger linguistic units. Precisely, Pilehvar and Navigli (2015) propose an unified method to compute the semantic similarity between items from multiple linguistic levels. On the other hand, Saif et al. (2014) have carried out a study on the impact of the incompleteness of some linguistic resources

Well-founded IC models	Definition
CondProbHypo	$IC_{CPHypo}(c_i) = -log_2(p_{Hypo}(c_i))$ $p_{Hypo}(c_i\|c_j) = \frac{\|Hypo(c_i)\|+1}{\sum_{\forall c_k \mid c_j \in LA(c_k)} (\|Hypo(c_k)\|+1)}$
CondProbUniform	$IC_{CPUni}(c_i) = -log_2(p_{Uniform}(c_i))$ $p_{Uniform}(c_i\|c_j) = \frac{1}{\|children(c_j)\|}$
CondProbLeaves	$IC_{CPLea}(c_i) = -log_2(p_{Leaves}(c_i))$ $p_{Leaves}(c_i\|c_j) = \frac{\|Leaves(c_i)\|+1}{\sum_{\forall c_k \mid c_j \in LA(c_k)} (\|Leaves(c_k)\|+1)}$
CondProbLogistic	$IC_{CPLog}(c_i) = -log_2(p_{Log}(c_i))$ $p_{Log}(c_i\|c_j) = \varphi_l(x) \circ p_{Hypo}(c_i\|c_j)$ $\varphi_l(x:k) = \frac{1}{1+e^{-k\left(x-\frac{1}{2}\right)}},\ k^* = 8$
CondProbCosine	$IC_{CPCos}(c_i) = -log_2(p_{Cos}(c_i))$ $p_{Cos}(c_i\|c_j) = \varphi_c(x) \circ p_{Hypo}(c_i\|c_j)$ $\varphi_c(x) = 1 - cos\left(\frac{\pi}{2}x\right)$
CondProbCorpus	$IC_{CPCorpus}(c_i) = -log_2(p(c_i))$ $p(c_i) = \begin{cases} 1 & , c_i = \Gamma \\ \sum_{\forall c_j \in LA(c_i)} p(c_j)\, p_{corpus}(c_i\|c_j) & , c_i \neq \Gamma \end{cases}$ $p_{corpus}(c_i\|c_j) = \frac{max\{1, f(c_i)\}}{\sum_{\forall c_k \mid c_j \in LA(c_k)} max\{1, f(c_k)\}}$

Table 5: Our current family of well-founded IC models introduced by Lastra-Díaz and García-Serrano (2015a) and evaluated in this work. Hypo(c_i) and $Leaves(c_i)$ denote respectively the set of subsumed concepts and leaf concepts for any concept $c_i \in C$, without including the base concept c_i.

in Arabic, such as WordNet and Wikipedia, and into the performance of the ontology-based and gloss-based similarity measures. This latter work shows degradation of the performance from most ontology-based similarity measures, which call our attention to the problems of extrapolating the results based on English benchmarks and resources. Another interesting issue is the availability of a large word similarity benchmark based on WordNet that would also include instances of concepts and multiple-word terms, in the spirit of the TR9856 dataset introduced by Levy et al. (2015).

In summary, the mainstream of research into ontology-based similarity measures is still the proposal of new intrinsic IC models and IC-based measures, such as that proposed by Pirró and Euzenat (2010), Meng et al. (2014), Gao et al. (2015) and our aforementioned works. However, we also find in the literature some new corpus-based IC models such as that introduced by Harispe et al. (2015a), and some relevant non IC-based measures such as that proposed by Sánchez et al. (2012) and Hadj Taieb et al. (2014b). In addition, there are several strategies that could be explored in order to overcome the lexical coverage limitation of the ontologies, and the selection of a specific similarity measure for a particular application is still an open problem.

3.1 Related work on IC models

In another recent work by Lastra-Díaz and García-Serrano (2015a), we provide an in-depth review of the state of the art in IC models. For this reason, this section only provides a summary of the literature on IC models, including a review of the latest IC models published after our aforementioned work.

In Resnik (1995) and subsequently Resnik (1999), the author introduces the first IC model reported in the literature. The Resnik IC model relies on a frequency counting method of the occurrences of a concept and its subsumed concepts into a corpus, that is also described in detail by (Pedersen, 2013, p.34), who uses the Resnik method to build the WordNet-based frequency files used in our experiments, Pedersen (2008b). The Resnik frequency counting method does not take the word senses into account; however, Pedersen (2010) proves that the IC models derived from a non sense-tagged corpus perform better than the sense-tagged ones. In order to overcome the drawbacks of the corpus-based IC models, Seco et al. (2004) introduce the first intrinsic IC model reported in the literature, whose core idea is that the IC models can be computed using only taxonomical features, such as the hyponym set ratio. During the last decade, the development of intrinsic IC models has become one of the mainstreams of research in the area. Among the main intrinsic IC models proposed in the literature, we find the works in Zhou et al. (2008a), Sebti and Barfroush (2008), Blanchard et al. (2008), Sánchez et al. (2011), Sánchez and Batet (2012), Yuan et al. (2013), and Hadj Taieb et al. (2014a), as shown in table 4, as well as the IC models introduced by Lastra-Díaz and García-Serrano (2015a) that are shown in table 5.

Finally, we have four recent works on IC models introduced by Adhikari et al. (2015), Harispe et al. (2015a), Aouicha and Taieb (2015) and Ben Aouicha et al. (2016b). First, Harispe et al. (2015a) introduce a family of corpus-based IC models based on the Belief function theoretical framework which is encouraged by the observation that the occurrences of a concept not only impact the IC value of the more general ancestor concepts, the so-called ancestors, but should also im-

pact the IC value of the more specific concepts, the so-called descendants. Harispe et al. (2015a) propose three different corpus-based IC models based on an adaptation of the classic *belief* and *plausibility* functions in the Demster-Shafer theory (DST), and the *pignistic* function. Second, Adhikari et al. (2015) introduce a new intrinsic IC model which is encouraged by the lack of integration in the previous IC models of a large combination of taxonomical features in order to distinguish several structural differences between concepts not considered before. The Adhikari et al. (2015) IC model integrates the relative depth, hyponym structure, subsumed leaves count and subsumer set count. Aouicha and Taieb (2015) introduce a new intrinsic IC model specifically designed for the MeSH biomedical ontology which has not been evaluated in WordNet. And finally, Ben Aouicha et al. (2016b) introduce a new intrinsic IC model on WordNet which is based on a new quatification of the ancestor set of each base concept. has not been included in our experiments. Tables 4 and 5 show the set of IC models that is implemented and evaluated in our experiments. This latest set of IC models, together with the recent IC models proposed by Harispe et al. (2015a) and Aouicha and Taieb (2015), represent, to the best of our knowledge, all the intrinsic and corpus-based IC models reported in the literature. On the other hand, Blanchard et al. (2008) IC_g is evaluated herein for the first time in a word similarity benchmark.

4 The proposed refinement

In Lastra-Díaz and García-Serrano (2015a), we propose a general framework to design IC models based on different methods for the estimation of the conditional probability between child and parent concepts, and we introduce a new family of IC models based on it, the so-called *well-founded IC models* shown in table 5. Our *IC models* are computed into four steps: (a) estimation of the conditional probabilities $p(c_i|c_j)$; (b) building of a total ordering of the concept set; (c) recovery of the concept probabilities $p(c_i)$ by using the recursive formula in equation (3); and (d) recovery of the IC values from the concept probabilities $p(c_i)$.

In order to eliminate the two drawbacks detailed in section 1.2, we introduce two refinements into the family of well-founded IC models and derive nine new IC models. First, in order to solve the problem related to the two cognitive IC models, we define a subsequent *normalization step* in the recovery of the concept probabilities in step (c) above, such that the overall sum of the probability on the leaf concepts is always 1 for these cases. Second, in order to warrant that the IC models satisfy the *growing monotonicity axiom*, such that $\forall c_i \leq_C c_j \Rightarrow IC(c_i) \geq IC(c_j)$, we define a new method for recovering the final concept probabilities based on the definition of the probability $p(c_i)$ as the sum of the probabilities of the leaf concepts subsumed by the concept c_i, instead of the direct value returned by the recursive formula in equation (3). Thus, we define a subsequent *subsumed probability recovery* step in the probability recovery step (d) above. We note that this new definition of the concept probabilities as the probability of their subsumed leaves matches the axiomatic construction of a discrete probability space, as introduced by Lastra-Díaz and García-Serrano (2015a), or any book on the subject, such as Ash and Doléans-Dade (2000). The new method to compute the final probabilities $p(c_i)$ from the conditional probabilities $p(c_i|c_j)$ matches the previous method in our aforementioned work whenever the taxonomy is tree-like, but it produces a slightly different probability function on taxonomies with multiple inheritance. This latest refinement is a sufficient condition to satisfy the *growing monotonicity axiom* regardless of the conditional probability model or the type of base taxonomy.

Refinement 1. In order to satisfy the growing monotonicity axiom regardless of the type of taxonomy, we introduce the following changes into the algorithm used to build the well-founded IC models. First, we introduce the growing monotonicity axiom as a further axiom into the definition of a *well-founded IC model*. And second, in order to satisfy the new axiom (4) the concept probability is defined as the sum of the probability of its subsumed leaves, instead of the direct value obtained from the recursive formula in equation (3), as was done in our aforementioned work.

Refinement 2. In order to warrant that the sum of leaf concept probabilities is 1 for any cognitive IC model, such as the *CondProbLogistic* and *CondProbCosine* introduced in Lastra-Díaz and García-Serrano (2015a), it is necessary to normalize the overall sum of leaf probabilities to 1.

All new IC models share the same algebraic and computational structure, being computed into six steps: (1) estimation of the conditional probabilities; (2) building of a total ordering of the concepts within the taxonomy; (3) recovery of the concept probabilities $p(c_i)$ by using the recursive formula in equation (3); (4) *unit normalization* of the probability of the leaf nodes only for the IC models based on non-linear transformations of the conditional probability; (5) computation of each concept probability $p(c_i)$ as the overall sum of the probability of its subsumed leaves; and finally, (6) computation of the IC values from the concept probabilities. In this way, the new steps (4) and (5) above eliminate the two aforementioned drawbacks, but the four remaining steps are identical to the original algorithm 1 in our previous work.

The two refinements above lead us to the reformulation of the *algorithm 1* to build the well-founded IC models introduced by Lastra-Díaz and García-Serrano (2015a). The previous *algorithm 1* is substituted by the *new algorithm* to build the well-founded IC models, which is summarized in table 6. Unlike the previous *algorithm 1*, the *new algorithm* only uses the iterative top-down procedure defined by the recursive formula in equation (3) in order to compute the probability of the leaf nodes, not the probability of each concept as was done in our aforementioned work. We recall that the probability recovery algorithm defined by the top-down

formula in equation (3) warrants that the overall sum of the leaf probabilities is 1 if the conditional probabilities $p(c_i|c_j)$ are well-defined and satisfy the constraint in equation (1). This latter fact is formalized into the proposition 2 below.

The *New Algorithm* in table 6 works on any type of taxonomy, and satisfies all the structure axioms in definition 1. The algorithm includes the two modifications proposed above in order to eliminate the two drawbacks found in our previous method. Thus, the proposed algorithm completely closes the algebraic and computational definition of the family of well-founded IC models, and it should be used in the design of any new intrinsic IC model.

Definition 1 (refined well-founded IC model) *Given a taxonomy of concepts $\mathcal{C} = (C, \leq_C, \Gamma)$, and an IC model defined by the function $IC : C \to \mathbb{R}^+ \cup \{0\}$, we call it a refined well-founded IC model if it can be written as $IC(c) = -log_2(p(c))$ where $p(c)$ is a concept-valued function as defined in equation (4), and the functions $p(c_i|c_j)$ are the conditional probabilities between any child concept c_i and its parent concepts c_j, which satisfy the edge-based property as defined in equation (1).*

(1) *Edge-based axiom.* The sum of conditional probabilities $p(c_i|c_j)$ of the children nodes c_i on any parent c_j node must be equal to 1, as defined in equation (1), where $LA(c_i)$ denotes the set of lowest ancestors (direct parents) of any concept c_i.

$$\sum_{\forall c_i | c_j \in LA(c_i)} p(c_i|c_j) = 1 \tag{1}$$

(2) *Leaf node probability axiom.* The overall probability of the leaf concepts sums 1, as defined in equation (2), and they are computed using the iterative top-down algorithm defined by equation (3).

$$\sum_{c_k \in L_C} p(c_k) = 1 \tag{2}$$

$$\begin{aligned} p &: C \to [0,1] \subset \mathbb{R} \\ p(c_i) &= \begin{cases} 1 & , c_i = \Gamma \\ \sum\limits_{\forall c_j \in LA(c_i)} p(c_j)\, p(c_i|c_j) & , c_i \neq \Gamma \end{cases} \end{aligned} \tag{3}$$

(3) *Probability node axiom.* The probability $p(c_i)$ for each concept $c_i \in C$ must be equal to the sum of the probability of each subsumed leaf concept $c_k \in Leaves(c_i) = \{c_k \in C \mid c_k \leq_C c_i \wedge c_k \text{ is a leaf concept}\}$, as defined in equation (4).

$$p(c_i) = \sum_{c_k \in Leaves(c_i)} p(c_k) \tag{4}$$

(4) *Monotonicity.* $\forall c_i, c_j \in C,\ c_i \leq_C c_j \Rightarrow IC(c_i) \geq IC(c_j)$

The axioms (1), (2) and (3) above allow us to define a new family of well-founded intrinsic IC models based on the estimation of the conditional probabilities $p(c_i|c_j)$ for each edge of the taxonomy, as shown in table 7. The axiom (3) is a sufficient condition for the satisfaction of the axiom (4), thus, the new refined IC models satisfy the monotonicity axiom by design. We call the new family as *refined well-founded IC models* in order to distinguish it from our previous IC models, and to emphasize the use of the new algorithm in table 6. In proposition 1, we show that given a taxonomy $(C, \leq_C, \Gamma)$, the definition of the concept probabilities according to axiom (3) is a sufficient condition to get a well-founded probability space, which moreover matches the standard axiomatic construction of any discrete probability space. In addition, we show in proposition 2 that axioms (1) and (2) of a well-founded IC model are sufficient conditions to build a leaf-valued function $p : L_C \subset C \to [0,1]$ that satisfies axiom (2) above and the second premise of proposition 1. Thus, proposition 2 proves that any well-founded IC model induces a well-founded probability space on any base taxonomy, and the whole system is supported by the structures derived from the conditional probabilities. The proofs of both propositions are included in appendix B of Lastra-Díaz and García-Serrano (2015a).

Proposition 1 *Be a taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$ defined by a partially ordered set $(C, \leq_C)$ with a distinguished supreme element Γ, called the root, and L_C the set of leaves in C. If a set-valued positive function P is defined from the leaf-valued function p as follows:*

$$(1) \quad \begin{aligned} &P : 2^\Gamma \to [0,1] \\ &P(A) = \sum_{c_k \in L_C \cap A} p(c_k) \end{aligned}$$

$$(2) \quad \begin{aligned} &p : L_C \subset C \to [0,1] \\ &\sum_{c_k \in L_C} p(c_k) = 1 \end{aligned}$$

then the following facts are satisfied: (1) P is a probability measure, and (2) the triplet $\left(\Gamma, 2^\Gamma, P\right)$ is a probability space.

Proposition 2 *Let a taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$ and L_C be the set of leaves in C. Given a concept-valued function p defined by*

$$\begin{aligned} p &: C \to [0,1] \\ p(c_i) &= \begin{cases} 1 & , \text{if } c_i = \Gamma \\ \sum\limits_{\forall c_j \in LA_C(c_i)} p(c_i|c_j)\, p(c_j) & , \text{otherwise} \end{cases} \end{aligned}$$

then $P(L_C) = 1$, as given below:

$$P(L_C) = \sum_{c_k \in L_C} p(c_k) = 1$$

4.1 The new family of IC models

This section introduces eight new intrinsic IC models called *CondProbRefHyponyms*, *CondProbRefUniform*, *CondProbRefLeaves*, *CondProbRefLogistic*,

New probability and IC recovery algorithm	
Input:	a rooted taxonomy $C = (C, \leq_C, \Gamma)$
	(1) $p(c_i \vert c_j)$ for each child and parent concepts.
Output	(2) $p : C \rightarrow [0,1] \subset \mathbb{R}$
	(3) $IC : C \times C \rightarrow \mathbb{R}^+ \cup \{0\}$
1:	Compute the conditional probabilities $p(c_i \vert c_j)$.
2:	Build a queue Q with a total ordering of the taxonomy $(C, \leq_C, \Gamma)$, such that every concept is in a subsequent position to every one of its parent concepts
	Remark: top-down computation of the leaf node probabilities
3:	foreach $c_i \in Q$
4:	$p(c_i) = \begin{cases} 1 & \text{, if } c_i = \Gamma \\ \sum_{\forall c_j \in LA_C(c_i)} p(c_i \vert c_j)\, p(c_j) & \text{, otherwise} \end{cases}$
5:	end foreach
	Remark: normalization of the overall leaf node probability (only if the $p(c_i \vert c_j)$ values do not satisfy axiom 1)
6:	$overallLeavesProb = \sum_{c_k \in Leaves(\Gamma)} p(c_k)$
7:	foreach $c_i \in Leaves(\Gamma)$
8:	$p(c_i) = \frac{p(c_i)}{overallLeavesProb}$
9:	end foreach
	Remark: bottom-up computation of the node probabilities
	Remark: for the computation of the probability of each node, $Leaves(c_i)$ denotes the set of subsumed leaf concepts inclusive c_i.
10:	foreach $c_i \in Q$
11:	$p(c_i) = \sum_{c_k \in Leaves(c_i)} p(c_k)$
12:	$IC(c_i) = -log_2 p(c_i)$
13:	end foreach

Table 6: New algortihm for the computation of the refined well-founded IC models.

CondProbRefCosine, *CondProbLogisticLeaves*, *CondProbRefCosineLeaves* and *CondProbRefLeavesSubsumersRatio*, and a new corpus-based IC model called *CondProbRefCorpus*. From the latter list, the first five intrinsic IC models and the *CondProbRefCorpus* IC model are derived from the corresponding IC models introduced by Lastra-Díaz and García-Serrano (2015a) by using the *new algorithm* to compute the probability and IC values detailed in table 6. On the other hand, the new intrinsic IC models called *CondProbLogisticLeaves*, *CondProbRefCosineLeaves* and *CondProbRefLeavesSubsumersRatio* are based on three new methods to estimate the conditional probabilities $p(c_i \vert c_j)$. The *CondProbLogisticLeaves* and, *CondProbRefCosineLeaves* IC models combine the conditional probability function $p_{Leaves}(c_i \vert c_j)$ with two different cognitive-based non-linear similarity functions previously introduced in our aforementioned work.

Because of the good performance exhibited by the Sánchez et al. (2011) IC model in combination with our *coswJ&C* similarity measure, we propose the *CondProbRefLeavesSubsumersRatio* IC model which is a reformulation of the Sánchez et al. (2011) IC model based on the general framework proposed by the family of IC models introduced herein. This new IC model is based on the fact that the difference in IC values between child and parent concepts in a tree-like taxonomy matches the logarithm of the conditional probability $p(c_i \vert c_j)$. This latest observation inspired the family of IC-based similarity measures introduced by Lastra-Díaz and García-Serrano (2015b), and from it follows that the Sánchez et al. (2011) IC model can be reformulated as the ratio between child and parent concepts of the function $\sigma(x)$ in table 7. The function $\sigma(x)$ is called *Sánchez-Batet-Isern estimator*, because $\sigma(x)$ can be interpreted as a taxonomical estimator of the concept probabilities. Precisely, the *CondProbRefLeavesSubsumersRatio* IC model defines a well-defined probability space from the kernel function of the Sánchez et al. (2011) IC model, and this same strategy could be used in order to reformulate other IC models, or taxonomy-based conditional probability estimators, in the general framework proposed by our family of IC models.

Table 7 shows the definition of the new family of IC models. For the formulas in table 7, $Hypo(c_i)$ and $Leaves(c_i)$ denote respectively the set of subsumed concepts and subsumed leaf concepts for any concept $c_i \in C$, without including the base concept c_i. Unlike our previous work, each concept probability denoted by $p^*(c_i)$ is defined as the sum of the probability of the subsumed leaf nodes in equation (4), instead of the value directly obtained from the top-down formula in equation (3). The probability values $p(c_i)$ of the non-leaf concepts that are obtained from the top-down formula in equa-

New IC models in this work	Definition
CondProbRefHyponym	$IC_{CPRefHypo}(c_i) = -log_2\left(p^*_{Hypo}(c_i)\right)$ $p_{Hypo}(c_i\|c_j) = \frac{\|Hypo(c_i)\|+1}{\sum_{\forall c_k \mid c_j \in LA(c_k)} (\|Hypo(c_k)\|+1)}$
CondProbRefUniform	$IC_{CPRefUni}(c_i) = -log_2\left(p^*_{Uniform}(c_i)\right)$ $p_{Uniform}(c_i\|c_j) = \frac{1}{\|children(c_j)\|}$
CondProbRefLeaves	$IC_{CPRefLea}(c_i) = -log_2\left(p^*_{Leaves}(c_i)\right)$ $p_{Leaves}(c_i\|c_j) = \frac{\|Leaves(c_i)\|+1}{\sum_{\forall c_k \mid c_j \in LA(c_k)} (\|Leaves(c_k)\|+1)}$
CondProbRefLogistic	$IC_{CPRefLog}(c_i) = -log_2\left(p^*_{Log}(c_i)\right)$ $p_{Log}(c_i\|c_j) = \varphi_l(x) \circ p_{Hypo}(c_i\|c_j)$ $\varphi_l(x:k) = \frac{1}{1+e^{-k\left(x-\frac{1}{2}\right)}}, \quad k^* = 8$
CondProbRefCosine	$IC_{CPRefCos}(c_i) = -log_2\left(p^*_{Cos}(c_i)\right)$ $p_{Cos}(c_i\|c_j) = \varphi_c(x) \circ p_{Hypo}(c_i\|c_j)$ $\varphi_c(x) = 1 - cos\left(\frac{\pi}{2}x\right)$
CondProbRefCorpus	$IC_{CPRefCorpus}(c_i) = -log_2\left(p^*(c_i)\right)$ $p_{corpus}(c_i\|c_j) = \frac{max\{1,f(c_i)\}}{\sum_{\forall c_k \mid c_j \in LA(c_k)} max\{1,f(c_k)\}}$
CondProbRefLogisticLeaves	$IC_{CPRefLogLeaves}(c_i) = -log_2\left(p^*_{LogLeaves}(c_i)\right)$ $p_{LogLeaves}(c_i\|c_j) = \varphi_l(x) \circ p_{Leaves}(c_i\|c_j)$ $\varphi_l(x:k) = \frac{1}{1+e^{-k\left(x-\frac{1}{2}\right)}}, \quad k^* = 8$
CondProbRefCosineLeaves	$IC_{CPRefCos}(c_i) = -log_2\left(p^*_{CosLeaves}(c_i)\right)$ $p_{CosLeaves}(c_i\|c_j) = \varphi_c(x) \circ p_{Leaves}(c_i\|c_j)$ $\varphi_c(x) = 1 - cos\left(\frac{\pi}{2}x\right)$
CondProbRefLeavesSubsumersRatio	$IC_{CPRefLeaSubRat}(c_i) = -log_2\left(p^*_{LeaSubRat}(c_i)\right)$ $p_{LeaSubRat}(c_i\|c_j) = \frac{\frac{\sigma(c_i)}{\sigma(c_j)}}{\sum_{\forall c_k \mid c_j \in LA(c_k)} \frac{\sigma(c_i)}{\sigma(c_j)}}$ $\sigma(c) = \frac{\|Leaves(c)\|}{\|subsumers(c)\|} + 1$

Table 7: New set of IC models proposed into the family of well-founded IC models. Unlike our previous work, each concept probability denoted by $p^*(c_i)$ is defined as the sum of the probability of the subsumed leaf nodes, instead of the value directly obtained from the recursive formula in equation (3). The new IC models are computed using the new algorithm detailed in Table 5. Hypo(c_i) and $Leaves(c_i)$ denote respectively the set of subsumed concepts and leaf concepts for any concept $c_i \in C$, without including the base concept c_i.

tion (3) are only temporary values whose aim is to obtain the estimated probability value of each leaf concept. The new IC models are computed using the new algorithm detailed in table 6. The *CondProbRefLogistic*, *CondProbRefCosine*, *CondProbLogisticLeaves* and *CondProbRefCosineLeaves* IC models do not satisfy the edge-based axiom defined by equation (1) in definition 1 because of they integrate a non-linear monotone transformation in their definition that prevents it, thus, the weights of the taxonomy used with the *coswJ&C* similarity measure in table 2 are set to $|IC(c_i) - IC(c_j)|$ instead of $-log_2(p(c_i|c_j))$.

5 Evaluation

The goals of the experiments described in this section are as follows: (1) the experimental evaluation of the proposed IC models and their comparison with the state-of-the-art methods; (2) a new experimental study onto the state of the art in ontology-based similarity measures; (3) a detailed statistical significance analysis of the similarity measures and IC models; (4) the replication of previously reported methods and results; (5) a new comparison between intrinsic and corpus-based IC models; (6) a study into the impact of the IC models on the IC-based similarity measures; (7) a comparison of the computational cost of the ontology-based similarity measures; (8) a new confirmation of the findings in our previous aforementioned works on the refuted outperformance of the intrinsic IC models over the corpus-based ones; and (9) a new confirmation of the achievements of the family of intrinsic IC models and IC-based similarity measures.

5.1 Methods evaluated herein

In order to compare the new family of IC models in table 7 with the state-of-the-art IC models, as well as providing a conclusive image of the state of the art of the problem, we implemented and evaluated all the IC models in tables 4, 5 and 7, as well as all the IC-based similarity measures in table 2 and the remaining ontology-based similarity measures shown in table 1. One IC model introduced by Blanchard et al. (2008) is evaluated herein for the first time. To the best of our knowledge, we

evaluate herein all WordNet-based intrinsic IC models reported in the literature, with the only exception of the IC model very recently introduced by Ben Aouicha et al. (2016b). Therefore, the experiments reported herein are the largest experimental survey of intrinsic IC models and ontology-based similarity measures reported up to date, which are based on a same code implementation.

For all the similarity measures and IC models, the depth is defined as the length of the shortest ascending path from each concept to the root. For the Zhou et al. IC model, the authors define the depth starting at 1 for the root concept. All methods have been implemented in a Java software library called HESML, which has been developed by the authors in order to replicate all methods evaluated herein. HESML was also used in our two aforementioned works on IC-based similarity measures and IC models, and it is going to be introduced and released in another forthcoming paper, Lastra-Díaz and García-Serrano (2016), together with a set of reproducible experiments and a replication dataset called *WNSimRep v1*.

In order to compare the intrinsic and corpus-based IC models, we use as baseline a corpus-based Resnik IC model based on the Wordnet-based frequency file called "ic-treebank-add1.dat" included in Pedersen (2008b), which was also used as a baseline in Lastra-Díaz and García-Serrano (2015a), having been the best performing corpus-based IC model in Lastra-Díaz and García-Serrano (2015b).

5.2 Experimental setup

We follow the same experimental setup defined by Lastra-Díaz and García-Serrano (2015a), including the same preprocessing steps, evaluation metrics, baselines, management of polysemic words and reporting of the results. In addition, we include for the first time a detailed pairwise statistical significance analysis between each pair of IC models and IC-based measures. We use the noun database of Wordnet 3.0, Miller (1995), and the five most significant word similarity benchmarks shown in table 8. For each word pair, we select the highest similarity value between the pairwise comparison of the sets of concepts evoked by each word.

Some preprocessing was necessary for the Agirre203 and SimLex-999 datasets to carry out the experiments. For the Agirre203 dataset, it was necessary to remove two word pairs containing verbs not present in the noun database of Wordnet 3.0, such as the pairs *(drink,eat)* and *(stock,live)*. In addition, it was also necessary to change the term "media" for "medium", and "children" for "child", because these terms do not appear directly in noun database. For this reason, we only used 201 nouns instead of 203, thus, this subset is called hereafter Agirre201. In the case of SimLex-999, it contains 666 nouns, but the word "august" is not included as synset in WordNet 3.0, thus, we only used 665 nouns from the SimLex-999 dataset, and this subset is called hereafter SimLex665. Finally, the MC30 dataset introduced by Miller and Charles (1991) is made up by 30 noun pairs; however, two word pairs are commonly ommited because of they were not included in previous versions of WordNet. For this reason, we use the MC28 dataset as defined at (Resnik, 1995, table 3) and (Li et al., 2003, p.875), together with the original human similarity judgements introduced by Rubenstein and Goodenough (1965). The datasets corresponding to the similarity benchmarks shown in table 8 are included in the HESML distribution.

Reference	Acronym	#wp	Description
Rubenstein and Goodenough (1965)	RG65	65	65 noun pairs ranging a similarity between 0 and 4.
Miller and Charles (1991)	MC28	28	Subset of RG65
Agirre et al. (2009)	Agirre201	201	Pure similarity subset of Finkelstein et al. (2002) with similarity in the range 0 to 10.
Pirró (2009)	$P\&S_{full}$	65	Modern replication of RG65
Hill et al. (2015)	SimLex665	665	Noun subset of SimLex-999 with similarity in the range 0 to 10.

Table 8: Word similarity benchmarks used in our experiments

5.3 Evaluation metrics

As evaluation metrics, we use the Pearson correlation factor, denoted by r in equation (5), and the Spearman rank correlation factor, denoted by ρ in equation (6). For a detailed review of the latter metrics, we refer the reader to (Lastra-Díaz and García-Serrano, 2015a, §5.3).

$$r = \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}} \quad (5)$$

$$\rho = 1-\frac{6\sum_{i=1}^{n}d_i^2}{n\left(n^2-1\right)},\quad d_i=(x_i-y_i) \quad (6)$$

In order to compare the performance of the IC models, we use the average Pearson and Spearman correlation values for each pair (IC model, IC-based similarity measure) on all datasets. The statistical significance of the results is evaluated by using the p-values resulting from the t-student test for the difference mean between the Spearman correlation values reported by each pair of IC models or IC-based similarity measures. The p-values are computed by using a one-sided t-student distribution on two paired sample sets. For the p-values between IC models, we use the vectors of the average Spearman correlation values over each IC-based similarity measure (rows in table 11) as a paired sample set, whilst for the similarity measures we use the vectors of Spearman correlation values of each similarity measure over all datasets (rows in table 12). Our null hypothesis,

denoted by H_0, is that the difference in the average performance between the compared IC models or IC-based measures is 0, whilst the alternative hypothesis, denoted by H_1, is that their average performance is different. For a 5% level of significance, it means that if the p-value is greater than 0.05, we must accept the null hypothesis, otherwise we can reject H_0 with an error probability of less than the p-value.

The Spearman rank correlation metric can represent better the use of the similarity measures in most rank-based selection tasks in NLP and IR, because it "provides an evaluation metric that is independent of these data-dependent transformations ", (Agirre et al., 2009, §6). In addition, most similarity measures are monotone transformations from previous measures. Therefore, a statistical significance analysis based on the Spearman correlation shows the intrinsic differences and similarities between methods in a more conclusive manner than an analysis based on the Pearson correlation. Likewise, in order to compare the IC-based similarity measures, we selected for each measure its best performing IC model according to the average Spearman correlation values shown in table 11.

5.4 Results obtained

Table 9 below shows the computational cost of each similarity measure on the MC28 dataset. The remaining data tables are included in the appendix next to the bibliography. Tables 10 and 11 show in each cell the average Pearson and Spearman correlation values respectively obtained in the evaluation of each IC model with any IC-based similarity measure on all datasets. Table 12 shows the Pearson and Spearman correlation values obtained by each ontology-based similarity measure on all datasets. In order to make the interpretation of the resulting p-values easier, tables 13 and 14 show a summary of the statistical significance analysis between the IC models and ontology-based similarity measures, whilst the raw p-values are shown in tables 25 and 26. Each row in tables 13 and 14 shows an 'x' whenever the method in the row header obtains a statistically significant higher performance than the method in the column header. Thus, the rows show the methods that are outperformed by each method on the left, whilst the columns show the methods that outperform each method at the top. Finally, tables 15 to 24 in the appendix show all raw data tables for the cross-evaluation of the IC models and IC-based similarity measures on all datasets.

Similarity measure	Overall (msec)	Avg (msec)	Ratio
Sánchez et al. (2012)	480	17.14	0.66
Pirró and Seco (2008)	696	24.86	0.96
Pirró and Euzenat (2010)	703	25.11	0.97
Garla and Brandt (2012)	715	25.54	0.98
Meng and Gu (2012)	716	25.57	0.99
Jiang and Conrath (1997)	722	25.79	0.99
Resnik (1995) (baseline)	**726**	**25.93**	**1.00**
Lin (1998)	728	26.00	1.00
Lastra-Díaz and García-Serrano (2015b), cosJ&C	735	26.25	1.01
Hadj Taieb et al. (2014b)	774	27.64	1.07
Al-Mubaid and Nguyen (2009)	38016	1357.71	52.36
Wu and Palmer (1994)	42514	1518.36	58.56
Gao et al. (2015)	44343	1583.68	61.08
Li et al. (2003), strategy 9	45201	1614.32	62.26
Meng et al. (2014)	48499	1732.11	66.80
Zhou et al. (2008b)	50343	1797.96	69.34
Pedersen et al. (2007)	53504	1910.86	73.70
Leacock and Chodorow (1998)	53921	1925.75	74.27
Li et al. (2003), strategy 4	54278	1938.50	74.76
Li et al. (2003), strategy 3	54607	1950.25	75.22
Rada et al. (1989)	56172	2006.14	77.37
Lastra-Díaz and García-Serrano (2015b), coswJ&C	172490	6160.36	237.59

Table 9: Overall running time and average time per word pair for each similarity measure in the MC28 dataset with the following PC setup: Windows 8.1 x64, Java 1.8, Intel Core i7-5570 @ 2.40 GHz, 8 Gb RAM. The rows are arranged in ascending order according to the running time reported for each similarity measure. All the similarity measures have been implemented and evaluated within a same software library developed by the authors. The last row shows the running time ratio as regard the baseline defined by the Resnik measure.

6 Discussion

6.1 Comparison of the IC models

Looking at tables 10 and 11, the following conclusions can be drawn: (1) the Seco et al. (2004) IC model obtains the highest average Pearson and Spearman correlation values on all datasets and IC-based similarity measures, as it is the best performing IC model on average; (2) a large set of IC models made up of the models introduced by Seco et al. (2004), Blanchard et al. (2008), Sánchez et al. (2011), Sánchez et al. (2012), Meng et al. (2012), Yuan et al. (2013) and Adhikari et al. (2015) obtain on average a higher Pearson and Spearman correlation values than the corpus-based IC model defined as baseline; (3) the new IC models called *CondProbRefHyponyms* and *CondProbRefCosine* obtain on average a higher Pearson and Spearman correlation values respectively than the baseline IC model, and the Zhou et al. (2008a) IC model also obtains on average a higher Spearman correlation value than the baseline IC model; (4) most of our family of well-founded IC models and the Hadj Taieb et al. (2014a) IC model obtain on average a lower Pearson and Spearman correlation values than the baseline IC model; and (5) the Hadj Taieb et al. (2014a) IC model obtains on average the lowest Pearson and Spearman correlation values among all IC models, and its average performance is much lower than the remaining IC models.

Tables 15 to 24 allow the following conclusions to be drawn: (1) the Sánchez et al. (2011) IC model obtains the highest Pearson correlation value with our *coswJ&C*

similarity measure in the RG65 dataset; (2) the Resnik IC model obtains the highest Pearson correlation value with the J&C similarity measure in the MC28 dataset; (3) our new *CondProbRefUniform* IC model obtains the highest Pearson correlation value with the FaITH similarity measure in the Agirre201 dataset; (4) the Yuan et al. (2013) IC model obtains the highest Pearson correlation value with the FaITH measure in the $P\&S_{full}$ dataset; and (5) the Seco et al. (2004) IC model obtains the highest Pearson correlation value with the Zhou et al. (2008b) similarity measure in the SimLex665 dataset. In addition, an analysis of the raw Spearman correlation values on all datasets allows the following conclusions to be drawn: (6) the Meng et al. (2012) IC model obtains the highest Spearman correlation value with our *coswJ&C* measure in the RG65 dataset; (7) the Resnik IC model obtains the highest Spearman correlation value with our *coswJ&C* measure in the MC28 dataset; (8) our new *CondProbRefUniform* IC model obtains the highest Spearman correlation value with the Lin (1998), FaITH and Meng and Gu (2012) similarity measures in the Agirre201 dataset; (9) the Sánchez et al. (2011) IC model obtains the highest Spearman correlation value with our *coswJ&C* similarity measure in the $P\&S_{full}$ dataset; and (10) the Yuan et al. (2013) IC model obtains the highest Spearman correlation value with the Zhou et al. (2008b) similarity measure in the SimLex665 dataset.

6.2 The statistical significance of the IC models

Table 13 allows the following conclusions to be drawn. First, the Seco et al. (2004) IC model obtains a statistically significant higher average Spearman correlation value than the remaining IC models with the only exception of the Sánchez et al. (2011) IC model. Second, Seco et al. (2004) and Sánchez et al. (2011) are the only IC models that are not outperformed in a statiscally significant manner by another IC model. Third, the Seco et al. (2004), Sánchez et al. (2011) and Yuan et al. (2013) IC models obtain a statistically significant higher average Spearman correlation value than the baseline defined by the corpus-based Resnik IC model, thus, this small set of state-of-the-art intrinsic IC models outperform the best performing corpus-based IC model, confirming the H3 hypothesis positively. Fourth, the Hadj Taieb et al. (2014a) IC model obtains a statistically significant lower average Spearman correlation than all of the IC models. Fifth, most of our intrinsic IC models obtain a statistically significant lower average Spearman correlation than the rest of the IC models, with the exception of the *CondProbHyponyms, CondProbCosine, CondProbRefHyponyms, CondProbRefLeaves* and *CondProbRefCosine* IC models. Sixth, the Zhou et al. (2008a), Meng et al. (2012) and Yuan et al. (2013) IC models only obtains a statistically significant lower average Spearman correlation than the Seco et al. (2004) IC model, whilst the Adhikari et al. (2015) IC model is only outperformed by another two IC models. Thus, the Zhou et al. (2008a), Meng et al. (2012), Yuan et al. (2013) and Adhikari et al. (2015) IC models follow the Seco et al. (2004) and Sánchez et al. (2011) IC models in terms of performance in the average Spearman correlation. However, looking at table 10, we see that the performance measured by the Pearson correlation of the Zhou et al. (2008a) IC model is much lower than the remaining IC models. And seventh, among the twenty-five IC models analyzed, the Resnik IC model defined as baseline obtains a statistically significant higher average Spearman correlation than other ten models, and it is statistically outperformed by only three intrinsic IC models, thus, there is no a statistically significant difference between most instrinsic IC models and the baseline, a fact that confirms the hypothesis H2 positively.

Finally, the hypothesis H8 behind the refinement and the new IC models introduced in this work is positively confirmed by the data obtained in our experiments. Looking at table 13, we see that the new IC models *CondProbRefUniform, CondProbRefLeaves, CondProbRefCosine* and *CondProbRefCorpus,* obtain a statistically significant higher average Spearman correlation than their corresponding non-refined IC models *CondProbUniform, CondProbLeaves, CondProbCosine* and *CondProbCorpus*. However, the *CondProbRefHyponyms* and *CondProbRefLogistic* IC models are not able to obtain a statistically significant higher performance than their corresponding models *CondProbHyponyms* and *CondProbLogistic*.

6.3 Comparison of the similarity measures

Table 12 shows that our *coswJ&C* similarity measure combined with the Sánchez et al. (2011) IC model obtains the highest Spearman correlation values in all datasets, with the only exception of SimLex665, the highest Pearson correlation values in the RG65 (0.8870) and MC28 (0.8710) datasets, as well as the highest overall average combined Pearson and Spearman correlation values (0.7708) shown in the last column and the highest overall average Spearman correlation value (0.7579). We point out that the highest Pearson correlation value (0.8809) in the MC28 dataset is obtained by the J&C similarity measure with the Resnik IC model, as shown in table 17, whilst the Seco et al. (2004) IC model is used for the overall comparison in table 12, because this latter IC model is the best performing IC model for the J&C measure in terms of the Spearman correlation.

Table 12 also shows that the Zhou et al. (2008b) similarity measure obtains the highest Pearson (0.6237) and Spearman (0.6101) correlation values in the SimLex665 dataset and the highest overall average Pearson correlation value (0.7859). In addition, the Zhou et al. (2008b) measure obtains the second best overall performance. The Hadj Taieb et al. (2014b) similarity measure obtains the highest Pearson correlation value (0.7123) in the Agirre201 dataset. The FaITH similarity measure introduced by Pirró and Euzenat (2010) obtains the highest Pearson correlation value (0.9082) in the $P\&S_{full}$ dataset when it is combined with the Yuan et al. (2013) model.

Table 12 shows that a small set of similarity measures obtain a higher overall performance than the baseline defined by J&C measure, as well as the Resnik and Lin similarity measures. This small set of outperforming measures is made up of our *coswJ&C* and *cosJ&C* similarity measures and the measures introduced by Zhou et al. (2008b), Pirró and Seco (2008), Hadj Taieb et al. (2014b) and Gao et al. (2015). In addition, a large set of ontology-based similarity measures obtain a higher average Pearson correlation value than the baseline defined by the J&C similarity measure.

The *coswJ&C* similarity measure, in combination with the Sánchez et al. (2011) IC model, obtains the best overall performance defined by the average of the Pearson and Spearman correlation values, as shown in last column of table 12. In addition, the *coswJ&C* similarity measure outperforms the remaining measures in the Spearman correlation metric. Looking at table 18, we can see another very meaningful and unexpected fact: *the coswJ&C similarity measure obtains the highest Spearman correlation value in the MC28 dataset with all the IC models, excluding the Hadj Taieb et al. (2014a) IC model.* We attribute the good performance of the *coswJ&C* similarity measure in the Spearman metric to the novel method for computing the distance between concepts that is defined by our distance $dis_{wJ\&C}$ introduced in Lastra-Díaz (2014) and Lastra-Díaz and García-Serrano (2015b), which defines an IC-based weighted graph as a generalization of the classic Jiang-Conrath distance. On the other hand, this $dis_{wJ\&C}$ measure requires the computation of the length of the shortest path on a non-uniform and real-valued weighted graph using the Dijkstra algorithm, whose computation time is longer than for the case in which only the edge count is required, as happens for the rest of the hybrid IC-based similarity measures shown in table 2. For this reason, the *coswJ&C* measure reports the highest computational cost in table 9, which is roughly three times greater than most hybrid IC-based similarity measures.

The data in table 9 allows the hypothesis H5 and the following conclusion introduced in Lastra-Díaz and García-Serrano (2015b) to be confirmed: despite the *coswJ&C* and Zhou et al. (2008b) similarity measures outperforming the remaining similarity measures on average, the computational cost and the performance of these measures, as well as the remaining hybrid IC-based similarity measures, prevent their use in practical applications. Thus, a practical option is to use our *cosJ&C* similarity measure, which obtains the third best overall performance, despite there being no statistical significant difference between it and the measures introduced by Pirró and Seco (2008) and Hadj Taieb et al. (2014b). Indeed, the general conclusion that we advance here is that the performance margin between the state-of-the-art ontology-based similarity measures is very narrow.

An interesting point is that the three similarity measures on top of table 12 are derived from the Jiang-Conrath distance. The *coswJ&C* similarity measure is a generalization of the Jiang-Conrath measure based on an IC-based weighted graph, whilst the Zhou et al. (2008b) similarity measure is a linear combination of it with the Leacock and Chodorow (1998) similarity measure. On the other hand, the *cosJ&C* similarity measures is a monotone transformation of the Jiang-Conrath distance. Thus, the measurement strategy introduced by Jiang and Conrath (1997) leads the state of the art of the problem.

6.4 The statistical significance of the similarity measures

Table 14 allows the following conclusions to be drawn: (1), our *coswJ&C* similarity measure and the measure introduced by Zhou et al. (2008b) obtain a statistically significant higher average Spearman correlation value than the baseline defined by the J&C measure, and they are the only measures that outperform the baseline; (2) our *coswJ&C* similarity measure, and the measures introduced by Zhou et al. (2008b) and Meng et al. (2014), are the only measures that are not outperformed by other measures in a statistically significant manner; (3) the Zhou et al. (2008b) similarity measure obtains a statistically significant higher average Spearman correlation value than all of the measures, with the only exception of the *coswJ&C* and Meng et al. (2014) similarity measures; (4) the Wu and Palmer (1994) similarity measure obtains a statistically significant lower average Spearman correlation value than all of the remaining measures; (5) our *coswJ&C* similarity measure and the measure introduced by Zhou et al. (2008b) obtain a statistically significant higher average Spearman correlation value than all of the classic IC-based measures, whilst our *cosJ&C* measure and the Garla and Brandt (2012) measure statistically outperform the Resnik and Lin similarity measures; and finally, (6) the Rada et al. (1989) similarity measure and all measures derived from it, such as the measures introduced by Leacock and Chodorow (1998) and Pedersen et al. (2007), together with the Al-Mubaid and Nguyen (2009) measure, are only outperformed in a statistically significant manner by our *coswJ&C* similarity measure, and the measures introduced by Zhou et al. (2008b) and Meng et al. (2014).

In summary, conclusions (1) and (2) above prove hypothesis H1 on the outperformance of the path-based similarity measures by a group of state-of-the-art IC-based similarity measures. Conclusion (5) above proves the hypothesis H4 on the outperformance of the classic IC-based similarity measures by a small set of state-of-the-art methods. On the other hand, the conclusion (6) above is very significant because it proves for the first time that *only this small set of state-of-the-art IC-based similarity measures have been able to obtain a statistically significant higher average Spearman correlation value than the family of path-based similarity measures.* If we reproduce the statistical significance analysis in table 14 using the average Pearson correlation as sample set, we could see that most IC-based similarity measures obtain a statistically significant higher average Pearson correlation than the path-based measures, a fact that endorses the common belief that the path-based similarity measures have been definitively outperformed by the

family of IC-based similarity measures. However, the results shown in table 14 reopen the debate. We argue that the lack of a statistically significant difference between the Garla and Brandt (2012) and Pedersen et al. (2007) similarity measures, and thus any other measure derived from Rada et al. (1989), is mainly responsible for the lack of a statistically significant difference in performance reported by Alonso and Contreras (2016) for the use of the two aforementioned measures in a biomedical IR task. The latter facts endorse our idea that research into the area should focus on the improvement in the performance based on the Spearman rank correlation, because this latter metric could predict the expected performance in applications based on similarity measures better.

We note other significant fact. Our *coswJ&C* similarity measure, and the measures introduced by Zhou et al. (2008b) and Meng et al. (2014), are all hybrid IC-based similarity measures that integrate an IC model with any path-based feature. Among the latter aforementioned measures, the *coswJ&C* similarity measure is the only one that defines a real IC-based weighted graph, whilst the other two measures integrate a pure edge-counting measure in their formulas. Our experimental results and the significance analysis show that the IC-based weighted distance on a taxonomy, as proposed by the *coswJ&C* similarity measure, is currently the best approach for maximizing the Spearman rank correlation value, thus, this type of taxonomical feature should be explored in future developments into ontology-based similarity measures, despite its high computational cost.

6.5 Impact of the IC models on the similarity measures

The last four rows in tables 10 and 11 show a set of statistics considering the Pearson and Spearman correlation values reported by each similarity measure (column) as a random variable evaluated on all IC models. These statistics allow the following conclusions to be drawn: (1) most IC-based similarity measures exhibit a moderate standard deviation in the Pearson and Spearman correlation values as regard the set of IC models; (2) most IC-based similarity measures in table 11 exhibit a peak ratio greater than 1.0 times their standard deviation, a fact that supports our H6 hypothesis which states that most IC-based similarity measures perform better with a specific IC model; and (3) the standard deviation of the Spearman correlation of the IC-based similarity measures as regards the IC models is statistically significant lower than the standard deviation of the Pearson correlation, a fact that is supported by a p-value of 0.0073 between both random sets. This latter fact means that the performance of the IC-based similarity measures as a function of the IC models is more stable in terms of the Spearman rank correlation than the Pearson metric.

We conclude that every similarity measure should be used with its best performing IC model in any practical application. However, there is no strong evidence confirming that the outperformance of a similarity measure in any word similarity benchmark can be extrapolated to other applications (see our discussion in section 3). Our most significant conclusion as regards the IC models is as follows: the two best performing and preferred IC models by most IC-based similarity measures, and thus, the most practical IC models, are those introduced by Sánchez et al. (2011) and Seco et al. (2004).

6.6 New state-of-the-art results

The new state-of-the-art in intrinsic IC models and intrinsic IC-based similarity measures is set by the Sánchez et al. (2011) IC model in combination with our *coswJ&C* similarity measure, and the Seco et al. (2004) IC model in combination with the Zhou et al. (2008b) similarity measure. Likewise, these two latter intrinsic IC-based similarity measures obtain a statistically significant higher performance than the remaining methods. Thus, the four aforementioned methods are convincing winners among the families of IC models and ontology-based similarity measures. The *coswJ&C* similarity measure obtains the highest average Spearman correlation value and the highest overall averaged Pearson-Spearman correlation value on all datasets, as well as the highest Spearman correlation value in four of the five datasets evaluated, and the highest Pearson correlation values in the RG65 and MC28 datasets. On the other hand, the Zhou et al. (2008b) similarity measure obtains the highest average Pearson correlation value on all datasets and the highest Spearman correlation value in the SimLex665 dataset.

The set of classic IC-based similarity measures, defined by the Resnik, Lin and Jiang-Conrath measures, have also been definitively outperformed in a statistically significant manner by a small set of IC-based similarity measures developed during the last decade, among which we find the similarity measures introduced by Zhou et al. (2008b) and the *coswJ&C* measure introduced by Lastra-Díaz and García-Serrano (2015b). In addition, the J&C similarity measure and its two monotone transformations, our *cosJ&C* measure and the Garla and Brandt (2012) similarity measure, obtain a statistically significant higher average Spearman correlation than the Resnik and Lin similarity measures, and the *cosJ&C* obtains a statistically significant average Pearson correlation value than the J&C similarity measure. However, we also prove that there is no a statistically significant difference between the two aforementioned pairs of outperforming IC-based similarity measures.

According to the results obtained, the two similarity measures with the best overall performance are the two hybrid IC-based similarity measures defined by the *coswJ&C* introduced by Lastra-Díaz and García-Serrano (2015b) and the Zhou et al. (2008b) measure. However, their computational cost prevents their practical use in comparison with other measures, such as the *cosJ&C* introduced by Lastra-Díaz and García-Serrano (2015b) and the Hadj Taieb et al. (2014b) measure. There is no statistically significant difference between these two latter measures. The *cosJ&C* measure obtains a higher Spearman correlation on average than the Hadj Taieb et al. (2014b) measure, whilst the Hadj Taieb et al. (2014b) measure obtains a higher Pearson cor-

relation on average than the previous one. Thus, the *cosJ&C* and Hadj Taieb et al. (2014b) measures are, statistically speaking, the best option from the aforementioned set of similarity measures with a practical computational cost.

6.7 Monotone transformations.

The Spearman rank correlation value is invariant to monotone transformations from any similarity measure, thus, its exhaustive evaluation for all the similarity measures and IC models has confirmed that a lot of similarity measures are monotone transformations of other classic measures, as well as the findings of other unknown cases. For instance, the Spearman correlation metric reported by the *FaITH* similarity measure introduced by Pirró and Euzenat (2010) reveals that it is a monotone transformation of the Lin measure like the measure introduced by Meng and Gu (2012). Indeed, there are many cases like these. For instance, the similarity measure introduced by Leacock and Chodorow (1998), the sim_{Path} measure of Pedersen et al. (2007), and the sim_{Li_s3} measure of Li et al. (2003), all which are monotone transformations of the Rada et al. (1989) measure, whilst the sim_{path_IC} measure of Garla and Brandt (2012) and the $sim_{cosJ\&C}$ measure introduced by Lastra-Díaz and García-Serrano (2015b) are monotone transformations of the J&C similarity measure as defined in table 2. We confirmed experimentally that in all of the aforementioned cases, the transformed measures preserve the Spearman correlation values obtained by their respective base measures, differing only in their Pearson correlation values. Table 3 shows a factorization of the latter similarity measures that proves the aforementioned monotonicity relationships.

As a consequence of the aforementioned monotonicity relationships, there is a reduced number of different strategies to estimate the degree of similarity using an ontology-based similarity measure, despite many similarity measures having been proposed in the literature. We argue that the monotonicity relationships between a large set of similarity measures are the main cause behind the lack of a statistically significant difference between most of the similarity measures evaluated herein. Thus, the research community should explore either new measurement methods or new similarity models in order to bring about significant progress in the state of the problem. On the other hand, the results obtained by the measures introduced by Meng and Gu (2012), Garla and Brandt (2012), Pirró and Euzenat (2010) and Lastra-Díaz and García-Serrano (2015b), prove that a proper scaling and normalization of the similarity measures is a good strategy to improve the Pearson correlation metric slightly. Therefore, the research should focus on the search for a significant improvement in the Spearman correlation metric, which is also closely related to the measurement strategy and similarity model used.

6.8 Computational complexity

Table 9 compares the running time of each similarity measure in the evaluation of the MC28 dataset. The feature-based measure of Sánchez et al. (2012) obtains the lowest running time, making it the fastest among all of the measures. As we expected from an analysis of their definitions, all non hybrid IC-based similarity measures obtain a running time that is almost identical to that reported by the Resnik measure defined as baseline. The small differences are only attributable to the activity of the operating system during the experiments, because these IC-based similarity measures share the same IC-based factors. On the other hand, the hybrid IC-based similarity measures exhibit a running time of between 52 and 237 times greater than the baseline, making our *coswJ&C* similarity measure the slowest among all of the measures. Thus, the computational complexity of the hybrid IC-based measures is roughly two orders of magnitude greater than the complexity of the remaining IC-based similarity measures. Despite all hybrid IC-based similarity measures using the same implementation of the Dijkstra algorithm in our software library, our *coswJ&C* similarity measure requires the measurement of the length of the shortest path between concepts on a non-uniform and real-valued weighted graph, whilst the rest of the hybrid IC-based similarity measures only require the edge count to be obtained, thus, the Dijsktra algorithm is much faster in this latter case.

6.9 Confirming our hypotheses

The hypotheses H1, H2, H3, H4, H5, H6 and H8 introduced in section 1.2 have been positively confirmed by the data obtained from our experiments, they having been answered in the discussion above. Finally, hypothesis H7 on the outperformance of the state-of-the-art IC-based similarity measures on the best corpus-based similarity measures in the SimLex666 dataset, is also confirmed by comparing the best Pearson and Spearman correlation values obtained by most IC-based similarity measures in tables 23 and 24, with the results for these metrics reported for the best corpus-based method in the SimLex dataset (Pearson=0.599, Spearman=0.591), as reported in a recent benchmark by Banjade et al. (2015).

6.10 Contradictory results

We obtained several contradictory results in our experiments, confirming the same findings reported in our aforementioned works, as well as other new ones. For instance, Meng and Gu (2012) and Meng et al. (2014) report Pearson correlation values of 0.8804 and 0.8817 respectively with the Seco et al. (2004) IC model in the RG65 dataset, whlist we obtained 0.8596 and 0.8486 respectively. Gao et al. (2015) report a Pearson correlation value of 0.885 for their similarity measure in the RG65 dataset with an unknown corpus-based IC model, whilst we obtained 0.87098 herein. Adhikari et al. (2015) report the following Pearson correlation values of 0.86, 0.86 and 0.84 for their IC model in the MC30 dataset with the Resnik, Lin and Jiang-Conrath similarity measure respectively, whilst we obtained 0.8211, 0.8410 and 0.8331 in the MC28 dataset. These facts confirm the reproducibility problems in the area. Thus, we invite the

research community to reproduce the methods and experiments reported in the literature in order to confirm or refute the results reported herein.

7 Conclusions and future work

We have introduced a refinement of our recent family of well-founded Information Content models, eight new intrinsic IC models and one new corpus-based IC model and a very detailed experimental survey on WordNet. We have proven that the proposed refinement improves the performance of our family of well-founded IC models, and six of our new IC models obtain rivaling results as regard the state-of-the-art intrinsic IC models, making the new *CondProbRefHyponyms* and *CondProbRefCosine* IC models our best performing IC models.

The Seco et al. (2004) and Sánchez et al. (2011) IC models set the state of the art for the IC models, and the Seco et al. (2004), Sánchez et al. (2011) and Yuan et al. (2013) IC models are the only intrinsic IC models that statistically outperfom the best performing corpus-based IC model used as baseline. However, we prove that there is no statistically significant difference between most intrinsic IC models and the corpus-based Resnik IC model defined as baseline. Therefore, the aforementioned set of intrinsic IC models can be considered as a practical alternative to the corpus-based ones, and they should be selected in accordance with the IC-based similarity measure used. On the other hand, the detailed experiment survey carried-out herein allows a very significant conclusion to be drawn: despite the research effort made during the last decade, the Seco et al. (2004) IC model is still the state of the art on average.

The new state-of-the-art in intrinsic IC models and intrinsic IC-based similarity measures is set by the Sánchez et al. (2011) IC model in combination with our *coswJ&C* similarity measure, and the Seco et al. (2004) IC model in combination with the Zhou et al. (2008b) similarity measure. The set of classic IC-based similarity measures, defined by the Resnik, Lin and Jiang-Conrath measures, have also been definitively outperformed in a statistically significant manner by a small set of IC-based similarity measures developed during the last decade, among which we find the similarity measures introduced by Zhou et al. (2008b) and the *coswJ&C* introduced by Lastra-Díaz and García-Serrano (2015b). In addition, the J&C similarity measure and its two monotone transformations, our *cosJ&C* measure and the Garla and Brandt (2012) similarity measure, statistically outperform the Resnik and Lin similarity measures, and the *cosJ&C* similarity measure obtains a statistically significant higher average Pearson correlation value than the J&C similarity measure. However, we also prove that there is no a statistically significant difference between the two aforementioned pairs of outperforming IC-based similarity measures.

Despite our *coswJ&C* similarity measure and the Zhou et al. (2008b) measure setting the state of the art of the problem, their computational cost prevent their practical use in comparison with other measures, such as the *cosJ&C* introduced by Lastra-Díaz and García-Serrano (2015b) and the Hadj Taieb et al. (2014b) measure. There is no a statiscally significant difference between the two latter aforementioned measures. Thus, the *cosJ&C* and Hadj Taieb et al. (2014b) measures are, statistically speaking, the best option from the aforementioned set of similarity measures with a practical computational cost.

We have proven that the state of the art in ontology-based similarity measures and concept similarity models is led by the family of IC-based measures, more specifically by the measures derived from the Jiang-Conrath similarity measure. In addition, we have made another significant finding. Contrary to the common belief among the research community, only a small set of state-of-the-art hybrid IC-based similarity measures derived from the J&C measure obtain a statistically significant higher average Spearman correlation value than the family of path-based similarity measures, a fact that explains some unexpected results in applications based on similarity measures reported in the literature, such as that reported by Alonso and Contreras (2016).

Finally, as forthcoming activities, we are going to introduce and releasing HESML in a forthcoming paper Lastra-Díaz and García-Serrano (2016), which is a new scalable Java software library of ontology-based semantic similarity measures and IC models. In addition, HESML will be released with a replication dataset called *WNSimRep v1*, as well as a set of reproducible experiments which allow automatically reproducing all the results reported in our two aforementioned works and herein. The aforementioned forthcoming paper is part of a novel innitiative on computational reproducibility recently introduced by Chirigati et al. (2016), whose pioneering work is introduced by Wolke et al. (2016) with the aim of aiding the exact replication of several dynamic resource allocation strategies in cloud data centers evaluated in another companion paper Wolke et al. (2015). Our reproducible experiments are based on ReproZip, which is a virtualization tool introduced by Chirigati et al. (2013b) and Chirigati et al. (2013a), whose aim is to warrant the exact replication of experimental results onto a different system from that originally used into their creation.

8 Acknowledgements

Ted Pedersen kindly answered all our questions and provided us the WordNet-based frequency files used to build the corpus-based IC models included in our experiments. Mohamed Hadj Taieb kindly offered us his total support in replicating their similarity measures exactly. Emmanuel Pothos, Vijay Garla, Abdulgabbar Saif, Jorge Martínez-Gil and Lubomir Stanchev kindly answered our questions about their works. Mark Hallett checked the proper use of English. To all of them, we would like to express our most sincere gratitude. Finally, we also express our gratitude to the anonymous reviewers by their remarks in order to improve the quality of this work. This work has been partially supported by the Spanish VOXPOPULI (TIN2013-47090-C3-1-P) Project.

9 Appendix

See summary tables 10, 11, 12, 13 and 14. All raw data resulting from the evaluation is shown in tables 15 to 26 next the bibliography.

References

Adhikari, A., Singh, S., Dutta, A., Dutta, B., Nov. 2015. A novel information theoretic approach for finding semantic similarity in WordNet. In: Proceedings of IEEE International Technical Conference (TENCON-2015). IEEE, Macau, China, pp. 1–6.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., Soroa, A., 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. NAACL '09. ACL, Stroudsburg, PA, USA, pp. 19–27.

Al-Mubaid, H., Nguyen, H. A., Jul. 2009. Measuring Semantic Similarity Between Biomedical Concepts Within Multiple Ontologies. IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews: a publication of the IEEE Systems, Man, and Cybernetics Society 39 (4), 389–398.

Alonso, I., Contreras, D., Feb. 2016. Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: an UMLS approach. Expert Systems with Applications 44, 386–399.

Alvarez, M. A., Lim, S., Sep. 2007. A Graph Modeling of Semantic Similarity between Words. In: Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007). IEEE Computer Society, Irvine, California, USA, pp. 355–362.

Aouicha, M. B., Taieb, M. A. H., 17 Dec. 2015. Computing semantic similarity between biomedical concepts using new information content approach. Journal of Biomedical Informatics.

Ash, R. B., Doléans-Dade, C. A., 2000. Probability & Measure Theory, 2nd Edition. Academic Press.

Banjade, R., Maharjan, N., Niraula, N. B., Rus, V., Gautam, D., 14 Apr. 2015. Lemon and Tea Are Not Similar: Measuring Word-to-Word Similarity by Combining Different Methods. In: Gelbukh, A. (Ed.), Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing). Vol. 9041 of LNCS. Springer, Cayro, Egypt, pp. 335–346.

Batet, M., Harispe, S., Ranwez, S., Sánchez, D., Ranwez, V., 1 Nov. 2014. An information theoretic approach to improve semantic similarity assessments across multiple ontologies. Information sciences 283, 197–210.

Batet, M., Sánchez, D., Valls, A., Feb. 2011. An ontology-based measure to compute semantic similarity in biomedicine. Journal of Biomedical Informatics 44 (1), 118–125.

Ben Aouicha, M., Hadj Taieb, M. A., Ezzeddine, M., Apr. 2016a. Derivation of "is a" taxonomy from Wikipedia Category Graph. Engineering Applications of Artificial intelligence 50, 265–286.

Ben Aouicha, M., Taieb, M. A. H., Ben Hamadou, A., 28 Mar. 2016b. Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. Applied Intelligence, 1–37.

Blanchard, E., Harzallah, M., Kuntz, P., 2008. A generic framework for comparing semantic similarities on a subsumption hierarchy. In: Ghallab, M., Spyropoulos, C. D., Fakotakis, N., Avouris, N. (Eds.), Proceedings of the ECAI. Vol. 178 of Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 20–24.

Budanitsky, A., Hirst, G., 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In: Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics. Vol. 2. ACL, Pittsburgh, PA, pp. 29–34.

Budanitsky, A., Hirst, G., Mar. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics 32 (1), 13–47.

Busemeyer, J. R., Bruza, P. D., 2012. Quantum models of cognition and decision. Cambridge University Press.

Chaves-González, J. M., Martínez-Gil, J., Jan. 2013. Evolutionary algorithm based on different semantic similarity functions for synonym recognition in the biomedical domain. Knowledge-Based Systems 37, 62–69.

Chiang, J.-H., Ho, S.-H., Wang, W.-H., Oct. 2008. Similar genes discovery system (SGDS): Application for predicting possible pathways by using GO semantic similarity measure. Expert Systems with Applications 35 (3), 1115–1121.

Chirigati, F., Capone, R., Rampin, R., Freire, J., Shasha, D., Mar. 2016. A collaborative approach to computational reproducibility. Information Systems 59, 95–97.

Chirigati, F., Shasha, D., Freire, J., 2013a. Packing Experiments for Sharing and Publication. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. SIGMOD '13. ACM, New York, NY, USA, pp. 977–980.

Chirigati, F., Shasha, D., Freire, J., 2013b. Reprozip: Using provenance to support computational reproducibility. In: Presented as part of the 5th USENIX Workshop on the Theory and Practice of Provenance. usenix.org.

Couto, F. M., Pinto, H. S., Oct. 2013. The next generation of similarity measures that fully explore the semantics in biomedical ontologies. Journal of Bioinformatics and Computational Biology 11 (5), 1371001.

Couto, F. M., Silva, M. J., Coutinho, P. M., Apr. 2007. Measuring semantic similarity between Gene Ontology terms. Data & Knowledge Engineering 61 (1), 137–152.

Cross, V., Hu, X., 2011. Using Semantic Similarity in Ontology Alignment. In: Proceedings of the Sixth International Workshop on Ontology Matching (OM), 10th Int. Semantic Web Conference (ISWC 2011). Bonn Germany, pp. 61–72.

Dagher, G. G., Fung, B. C. M., Jul. 2013. Subject-based semantic document clustering for digital forensic investigations. Data & Knowledge Engineering 86, 224–241.

Fernando, S., Stevenson, M., 2008. A semantic similarity approach to paraphrase detection. In: Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics. Oxford, UK, pp. 45–52.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin., E., 1 Jan. 2002. Placing search in context: the concept revisited. ACM Transactions on Information Systems (TOIS) 20 (1), 116–131.

Fokkens, A., Van Erp, M., Postma, M., Pedersen, T., Vossen, P., Freire, N., 4 Aug. 2013. Offspring from Reproduction Problems: What Replication Failure Teaches Us. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. ACL, Sofia, Bulgaria, pp. 1691–1701.

Gabrilovich, E., Markovitch, S., 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07). Vol. 7. Morgan Kaufmann Publishers Inc., Hyderabad, India, pp. 1606–1611.

Gao, J. B., Zhang, B. W., Chen, X. H., Mar. 2015. A WordNet-based semantic similarity measurement combining edge-counting and information content theory. Engineering Applications of Artificial Intelligence 39, 80–88.

Garla, V. N., Brandt, C., 10 Oct. 2012. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC bioinformatics 13:261.

Hadj Taieb, M. A., Ben Aouicha, M., Ben Hamadou, A., 1 Nov. 2014a. A new semantic relatedness measurement using WordNet features. Knowledge and Information Systems 41 (2), 467–497.

Hadj Taieb, M. A., Ben Aouicha, M., Ben Hamadou, A., Nov. 2014b. Ontology-based approach for measuring semantic similarity. Engineering Applications of Artificial Intelligence 36, 238–261.

Hadj Taieb, M. A., Ben Aouicha, M., Bourouis, Y., 22 Jun. 2015. Fm3s: Features-based measure of sentences semantic similarity. In: Onieva, E., Santos, I., Osaba, E., Quintián, H., Corchado, E. (Eds.), Proceedings of the 10th International Conference on Hybrid Artificial Intelligent Systems (HAIS 2015). Vol. 9121 of LNCS. Springer, Bilbao, Spain, pp. 515–529.

Harispe, S., Imoussaten, A., Trousset, F., Montmain, J., Aug. 2015a. On the consideration of a bring-to-mind model for computing the Information Content of concepts defined into ontologies. In: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2015). IEEE, Istambul, Turkey, pp. 1–8.

Harispe, S., Ranwez, S., Janaqi, S., Montmain, J., May 2015b. Semantic Similarity from Natural Language and Ontology Analysis. Vol. 8 of Synthesis Lectures on Human Language Technologies. Morgan & Claypool publishing.

Hill, F., Reichart, R., Korhonen, A., 2015. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. Computational Linguistics 41 (4), 665–695.

Hirst, G., St-Onge, D., 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum, C. (Ed.), WordNet: An electronic lexical database. Massachusetts Institute of Technology, pp. 305–332.

Hughes, T., Ramage, D., 28 Jun. 2007. Lexical Semantic Relatedness with Random Graph Walks. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). ACL, Prague, Czech Republic, pp. 581–589.

Jeong, B., Lee, D., Cho, H., Lee, J., Apr. 2008. A novel method for measuring semantic similarity for xml schema matching. Expert Systems with Applications 34 (3), 1651–1658.

Jiang, J. J., Conrath, D. W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference Research on Computational Linguistics (ROCLING X). pp. 19–33.

Jurgens, D., Pilehvar, M. T., Navigli, R., Oct. 2015. Cross level semantic similarity: an evaluation framework for universal measures of similarity. Language Resources and Evaluation, 1–29.

Lastra-Díaz, J. J., 29 Sep. 2014. Intrinsic Semantic Spaces for the representation of documents and semantic annotated data. Department of Computer Languages and Systems. Universidad Nacional de Educación a Distancia (UNED). http://e-spacio.uned.es/fez/view/bibliuned:master-ETSIInformatica-LSI-Jlastra.

Lastra Díaz, J. J., García Serrano, A., 19 Dec. 2014. System and method for the indexing and retrieval

of semantically annotated data using an ontology-based information retrieval model. United States Patent and Trademark Office (USPTO) application US14/576,679.

Lastra-Díaz, J. J., García-Serrano, A., Nov. 2015a. A new family of information content models with an experimental survey on WordNet. Knowledge-Based Systems 89, 509–526.

Lastra-Díaz, J. J., García-Serrano, A., Nov. 2015b. A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. Engineering Applications of Artificial Intelligence Journal 46, 140–153.

Lastra-Díaz, J. J., García-Serrano, A., 2016. HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. To appear in Information Systems journal.

Le, M., Fokkens, A., 7 Sep. 2015. Taxonomy Beats Corpus in Similarity Identification, but Does It Matter? In: Angelova, G., Bontcheva, K., Mitkov, R. (Eds.), Proceedings of International Conference on Recent Advances in Natural Language Processing. Hissar, Bulgaria, pp. 346–354.

Leacock, C., Chodorow, M., 1998. Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (Ed.), WordNet: An electronic lexical database. MIT Press, Ch. 11, pp. 265–283.

Lee, M. C., May 2011. A novel sentence similarity measure for semantic-based expert systems. Expert Systems with Applications 38 (5), 6392–6399.

Levy, R., Ein-Dor, L., Hummel, S., Rinott, R., Slonim, N., 26 Jul. 2015. TR9856: A Multi-word Term Relatedness Benchmark. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers). ACL, Beijing, China, pp. 419–424.

Li, P., Wang, H., Zhu, K. Q., Wang, Z., Hu, X.-G., Wu, X., 2015. A Large Probabilistic Semantic Network based Approach to Compute Term Similarity. IEEE Transactions on Knowledge and Data Engineering 27 (10), 2604–2617.

Li, Y., Bandar, Z. A., McLean, D., 2003. An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering 15 (4), 871–882.

Lin, D., 1998. An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning. Vol. 98. Madison, WI, pp. 296–304.

Liu, M., Shen, W., Hao, Q., Yan, J., Dec. 2009. An weighted ontology-based semantic similarity algorithm for web service. Expert Systems with Applications 36 (10), 12480–12490.

Martínez, S., Sánchez, D., Valls, A., 27 Oct. 2010. Ontology-based anonymization of categorical values. In: Modeling Decisions for Artificial Intelligence. Vol. 6408 of LNCS. Springer Berlin Heidelberg, pp. 243–254.

Martinez-Gil, J., Dec. 2016. CoTO: A Novel Approach for Fuzzy Aggregation of Semantic Similarity Measures. Cognitive Systems Research 40, 8–17.

McInnes, B. T., Pedersen, T., Dec. 2013. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. Journal of biomedical informatics 46 (6), 1116–1124.

Meijer, K., Frasincar, F., Hogenboom, F., Jun. 2014. A semantic approach for extracting domain taxonomies from text. Decision Support Systems 62, 78–93.

Meng, L., Gu, J., 2012. A New Model for Measuring Word Sense Similarity in WordNet. In: Proceedings of the 4th International Conference on Advanced Communication and Networking, ASTL. Vol. 14. pp. 18–23.

Meng, L., Gu, J., Zhou, Z., Sep. 2012. A new model of information content based on concept's topology for measuring semantic similarity in WordNet. International Journal of Grid and Distributed Computing 5 (3), 81–93.

Meng, L., Huang, R., Gu, J., Jun. 2014. Measuring Semantic Similarity of Word Pairs Using Path and Information Content. International Journal of Future Generation Communication & Networking 7 (3), 183–194.

Mihalcea, R., Corley, C., Strapparava, C., 2006. Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 1. AAAI Press, pp. 775–780.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. Q. (Eds.), Advances in Neural Information Processing Systems 26 (NIPS 2013). NIPS Foundation, Inc., pp. 3111–3119.

Miller, G. A., 1995. WordNet: A Lexical Database for English. Communications of the ACM 38 (11), 39–41.

Miller, G. A., Charles, W. G., 1991. Contextual correlates of semantic similarity. Language and cognitive processes 6 (1), 1–28.

Mohammad, S., Hirst, G., 2006. Distributional Measures of Concept-distance: A Task-oriented Evaluation. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. EMNLP '06. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 35–43.

Mohammad, S. M., Hirst, G., 8 Mar. 2012. Distributional Measures of Semantic Distance: A Survey. arXiv:1203.1858.

Montani, S., Leonardi, G., Quaglini, S., Cavallini, A., Micieli, G., 1 Jun. 2015. A knowledge-intensive approach to process similarity calculation. Expert Systems with Applications 42 (9), 4207–4215.

Oliva, J., Serrano, J. I., del Castillo, M. D., Iglesias, A., Apr. 2011. SyMSS: A syntax-based measure for short-text semantic similarity. Data & Knowledge Engineering 70 (4), 390–405.

Patwardhan, S., Banerjee, S., Pedersen, T., Feb. 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Gelbukh, A. (Ed.), Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2003). Vol. 2588 of LNCS. Springer, Mexico D.F., pp. 241–257.

Patwardhan, S., Pedersen, T., 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together. Vol. 1501. Trento, Italy, pp. 1–8.

Pedersen, T., 2008a. Empiricism Is Not a Matter of Faith. Computational Linguistics 34 (3), 465–470.

Pedersen, T., 2008b. WordNet-InfoContent-3.0.tar dataset repository. https://www.researchgate.net/publication/273885902_WordNet-infoContent-3.0.tar.

Pedersen, T., 2010. Information Content Measures of Semantic Similarity Perform Better Without Sense-tagged Text. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. HLT '10. ACL, Stroudsburg, PA, USA, pp. 329–332.

Pedersen, T., 25 Nov. 2013. Measuring the Similarity and Relatedness of Concepts: a MICAI 2013 Tutorial. Invited talk in the 12th Mexican International Conference on Artificial Intelligence (MICAI 2013). http://dx.doi.org/10.13140/RG.2.1.3025.6164.

Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., Chute, C. G., Jun. 2007. Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics 40 (3), 288–299.

Pennington, J., Socher, R., Manning, C. D., 2014. Glove: Global vectors for word representation. Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014) 12, 1532–1543.

Pesquita, C., Faria, D., Falcao, A. O., Lord, P., Couto, F. M., 2009. Semantic similarity in biomedical ontologies. PLoS computational biology 5 (7), e1000443.

Pilehvar, M. T., Navigli, R., Nov. 2015. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. Artificial Intelligence 228, 95–128.

Pirró, G., Nov. 2009. A semantic similarity metric combining features and intrinsic information content. Data & Knowledge Engineering 68 (11), 1289–1308.

Pirró, G., Euzenat, J., 7 Nov. 2010. A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In: Patel-Schneider, P. F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J. Z., Horrocks, I., Glimm, B. (Eds.), Proceedings of the 9th International Semantic Web Conference, ISWC 2010. Vol. 6496 of LNCS. Springer, Shangai, China, pp. 615–630.

Pirró, G., Seco, N., Jan. 2008. Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content. In: Meersman, R., Tari, Z. (Eds.), On the Move to Meaningful Internet Systems: OTM 2008. Vol. 5332 of LNCS. Springer, pp. 1271–1288.

Pirró, G., Talia, D., May 2010. Ufome: An ontology mapping system with strategy prediction capabilities. Data & Knowledge Engineering 69 (5), 444–471.

Pothos, E. M., Barque-Duran, A., Yearsley, J. M., Trueblood, J. S., Busemeyer, J. R., Hampton, J. A., 25 Feb. 2015. Progress and current challenges with the quantum similarity model. Frontiers in psychology 6, 205.

Pothos, E. M., Busemeyer, J. R., Trueblood, J. S., Jul. 2013. A quantum geometric model of similarity. Psychological review 120 (3), 679–696.

Pothos, E. M., Trueblood, J. S., Feb. 2015. Structured representations in a quantum probability model of similarity. Journal of Mathematical Psychology 64–65, 35–43.

Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics 19 (1), 17–30.

Ramage, D., Rafferty, A. N., Manning, C. D., 2009. Random Walks for Text Semantic Similarity. In: Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing. TextGraphs-4. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 23–31.

Resnik, P., 20 Aug. 1995. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI 1995). Vol. 1. Montreal, Canada, pp. 448–453.

Resnik, P., Jul. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11, 95–130.

Rubenstein, H., Goodenough, J. B., Oct. 1965. Contextual Correlates of Synonymy. Communications of the ACM 8 (10), 627–633.

Saif, A., Ab Aziz, M. J., Omar, N., 2014. Evaluating Knowledge-Based Semantic Measures on Arabic. International Journal on Communications Antenna and Propagation (IRECAP) 4 (5), 180–194.

Sánchez, D., Batet, M., Oct. 2011. Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. Journal of biomedical informatics 44 (5), 749–759.

Sánchez, D., Batet, M., 2012. A new model to compute the information content of concepts from taxonomic knowledge. International Journal on Semantic Web and Information Systems (IJSWIS) 8 (2), 34–50.

Sánchez, D., Batet, M., Isern, D., Mar. 2011. Ontology-based information content computation. Knowledge-Based Systems 24 (2), 297–303.

Sánchez, D., Batet, M., Isern, D., Valls, A., Jul. 2012. Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications 39 (9), 7718–7728.

Sanfilippo, A., Tratz, S., Gregory, M., Chappell, A., Whitney, P., Posse, C., Paulson, P., Baddeley, B., Hohimer, R., White, A., 7 Nov. 2005. Ontological annotation with wordnet. In: Proceedings of the 5th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot 2005) located at the 4rd International Semantic Web Conference (ISWC 2005). Galway, Ireland, pp. 27–36.

Sebti, A., Barfroush, A. A., Oct. 2008. A new word sense similarity measure in WordNet. In: Proceedings of the International Multiconference on Computer Science and Information Technology, IMCSIT 2008. IEEE, pp. 369–373.

Seco, N., Veale, T., Hayes, J., 2004. An intrinsic information content metric for semantic similarity in WordNet. In: López de Mántaras, R., Saitta, L. (Eds.), Proceedings of the 16th European Conference on Artificial Intelligence (ECAI). Vol. 16. IOS Press, Valencia, Spain, pp. 1089–1094.

Solé-Ribalta, A., Sánchez, D., Batet, M., Serratosa, F., Jan. 2014. Towards the estimation of feature-based semantic similarity using multiple ontologies. Knowledge-Based Systems 55, 101–113.

Song, W., Li, C. H., Park, S. C., Jul. 2009. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. Expert Systems with Applications 36 (5), 9095–9104.

Stanchev, L., 2 Jun. 2014. Creating a Similarity Graph from WordNet. In: Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS'14). Article No. 36. ACM.

Strube, M., Ponzetto, S. P., 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In: Proceedings of the AAAI Conference. Vol. 6. AAAI, pp. 1419–1424.

Suzuki, J., Nagata, M., Jul. 2015. A Unified Learning Framework of Skip-Grams and Global Vectors. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). ACL, Beijing, China, pp. 186–191.

Tversky, A., Jul. 1977. Features of similarity. Psychological Review 84 (4), 327–352.

Wang, Y., Zhou, Z., 20 Sep. 2009. Domain Ontology Generation Based on WordNet and Internet. In: Proceedings of the International Conference on Management and Service Science, 2009. MASS '09. IEEE, Wuhan, China, pp. 1–5.

Wolke, A., Bichler, M., Chirigati, F., Steeves, V., Jul. 2016. Reproducible experiments on dynamic resource allocation in cloud data centers. Information Systems 59, 98–101.

Wolke, A., Tsend-Ayush, B., Pfeiffer, C., Bichler, M., Aug. 2015. More than bin packing: Dynamic resource allocation strategies in cloud data centers. Information Systems 52, 83–95.

Wu, Z., Palmer, M., 1994. Verbs Semantics and Lexical Selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. ACL '94. ACL, Stroudsburg, PA, USA, pp. 133–138.

Yazdani, M., Popescu-Belis, A., Jan. 2013. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. Artificial Intelligence 194, 176–202.

Yeh, E., Ramage, D., Manning, C. D., Agirre, E., Soroa, A., 2009. WikiWalk: Random Walks on Wikipedia for Semantic Relatedness. In: Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing. TextGraphs-4. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 41–49.

Yuan, Q., Yu, Z., Wang, K., Dec. 2013. A New Model of Information Content for Measuring the Semantic Similarity between Concepts. In: Proceedings of the International Conference on Cloud Computing and Big Data (CloudCom-Asia 2013). IEEE Computer Society, pp. 141–146.

Zhou, Z., Wang, Y., Gu, J., 2008a. A new model of information content for semantic similarity in WordNet. In: Proc .of the Second International Conference on Future Generation Communication and Networking Symposia (FGCNS'08). Vol. 3. IEEE, pp. 85–89.

Zhou, Z., Wang, Y., Gu, J., Nov. 2008b. New model of semantic similarity measuring in WordNet. In: Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering (ISKE 2008). Vol. 1. IEEE, pp. 256–261.

Average - r	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)					
IC models	Resnik	Lin	$J\&C$	$P\&S$	FaITH	Meng_{12}	Garla	$cosJ\&C$	Li_{s9}	Zhou	Meng_{14}	Gao	$coswJ\&C$	Avg
Seco et al. (2004)	0.7385	0.7731	**0.7705**	**0.7784**	0.7743	0.7758	0.7762	0.7790	0.7486	**0.7859**	0.7373	0.7241	0.7727	0.7642
Adhikari et al. (2015)	0.7373	0.7714	0.7645	0.7717	0.7794	0.7792	0.7818	0.7762	0.7528	0.7744	0.7219	0.7241	0.7784	0.7626
Meng et al. (2012)	0.7367	0.7704	0.7632	0.7708	0.7794	0.7790	**0.7821**	0.7758	0.7539	0.7732	0.7185	0.7241	0.7784	0.7620
Yuan et al. (2013)	0.7372	**0.7744**	0.7613	0.7745	**0.7826**	**0.7830**	0.7788	0.7736	0.7511	0.7738	0.7187	0.7241	0.7695	0.7617
Sánchez et al. (2011)	**0.7459**	0.7612	0.7681	0.7174	0.7780	0.7752	0.6842	**0.7811**	0.7674	0.7702	0.7058	0.7733	**0.7836**	0.7547
Sánchez and Batet (2012)	0.7424	0.7727	0.7668	0.7408	0.7735	0.7753	0.6223	0.7752	0.7725	0.7694	0.7355	0.7756	0.7711	0.7533
Blanchard et al. (2008)	0.7371	0.7708	0.7660	0.7392	0.7708	0.7727	0.6232	0.7752	0.7737	0.7690	0.7372	0.7760	0.7709	0.7525
CPRefHyponyms	0.7325	0.7682	0.7651	0.7357	0.7686	0.7704	0.6247	0.7738	0.7748	0.7681	**0.7374**	0.7762	0.7727	0.7514
Resnik (1995)	0.7416	0.7690	0.7632	0.7404	0.7736	0.7742	0.6195	0.7718	0.7714	0.7660	0.7290	0.7768	0.7717	0.7514
CPRefCosine	0.7297	0.7689	0.7655	0.7363	0.7683	0.7704	0.6292	0.7729	0.7736	0.7689	0.7365	0.7757	0.7659	0.7509
CPRefLeaves	0.7319	0.7678	0.7633	0.7338	0.7689	0.7705	0.6244	0.7726	**0.7750**	0.7662	0.7369	0.7762	0.7719	0.7507
CPHyponyms	0.7330	0.7673	0.7659	0.7332	0.7688	0.7705	0.6236	0.7739	0.7717	0.7686	0.7300	0.7742	0.7727	0.7502
CPRefCosineLeaves	0.7301	0.7683	0.7626	0.7331	0.7689	0.7706	0.6279	0.7710	0.7738	0.7659	0.7359	0.7759	0.7654	0.7499
Zhou et al. (2008a)	0.7236	0.7418	0.7411	0.7497	0.7734	0.7676	0.7792	0.7645	0.7669	0.7694	0.6753	0.7241	0.7677	0.7496
CPRefLeaSubRatio	0.7352	0.7693	0.7617	0.7322	0.7708	0.7723	0.6120	0.7691	0.7747	0.7636	0.7363	0.7764	0.7708	0.7496
CPRefCorpus	0.7364	0.7646	0.7561	0.7324	0.7678	0.7686	0.6376	0.7645	0.7749	0.7597	0.7328	**0.7782**	0.7704	0.7495
CPCosine	0.7340	0.7687	0.7676	0.7344	0.7694	0.7712	0.5819	0.7750	0.7660	0.7692	0.7335	0.7698	0.7708	0.7470
CPLeaves	0.7326	0.7670	0.7641	0.7314	0.7691	0.7705	0.5457	0.7726	0.7718	0.7667	0.7297	0.7743	0.7719	0.7436
CPRefLogistic	0.7148	0.7550	0.7530	0.7221	0.7670	0.7659	0.6278	0.7657	0.7720	0.7561	0.7324	0.7656	0.7646	0.7432
CPRefLogisticLeaves	0.7152	0.7548	0.7529	0.7217	0.7677	0.7664	0.6273	0.7657	0.7716	0.7560	0.7313	0.7652	0.7649	0.7431
CPCorpus	0.7353	0.7630	0.7553	0.7288	0.7675	0.7681	0.5600	0.7641	0.7715	0.7589	0.7247	0.7766	0.7704	0.7419
CPLogistic	0.7142	0.7659	0.7296	0.7125	0.7714	0.7727	0.5985	0.7477	0.7699	0.7335	0.7181	0.7716	0.7573	0.7356
CPRefUniform	0.6610	0.7633	0.6704	0.6671	0.7742	0.7737	0.5982	0.6876	0.7635	0.6745	0.7067	0.7677	0.7334	0.7109
CPUniform	0.6135	0.7031	0.6376	0.6037	0.7362	0.7297	0.5877	0.6557	0.7632	0.6422	0.6538	0.7624	0.7334	0.6786
Hadj Taieb et al. (2014a)	0.4233	0.6765	0.3267	0.4196	0.6882	0.6857	0.4890	0.3278	0.7481	0.3311	0.6616	0.7549	0.3278	0.5277
Best column value	0.7459	0.7744	0.7705	0.7784	0.7826	0.7830	0.7821	0.7811	0.7750	**0.7859**	0.7374	0.7782	0.7836	0.7642
Average per column*	0.7246	0.7633	0.7515	0.7309	0.7704	0.7706	0.6481	0.7627	0.7678	0.7571	0.7231	0.7628	0.7675	
Std. deviation*	0.0290	0.0147	0.0317	0.0355	0.0085	0.0096	0.0740	0.0292	0.0082	0.0322	0.0205	0.0207	0.0117	
Peak ratio*	0.7348	0.7538	0.6004	1.3370	1.4430	1.2956	1.8116	0.6297	0.8847	0.8956	0.6938	0.7420	1.3759	

Table 10: Average on all datasets of the Pearson (r) correlation values for each pair (IC model, IC measure). The IC models in bold are the new methods introduced in this work. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score on all datasets for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column. The baseline is defined by the corpus-based Resnik IC model. The rows are arranged in descending order according to the average score in last column. (*) The Hadj Taieb et al. (2014a) IC model can be considered as an outlier, thus, we excluded it from the computation of these statistics. The peak ratio (pr) for a random variable X is defined by the equation $pr(X) = \frac{|max(X)-\overline{X}|}{\sigma(X)}$, $\sigma(X)$ being the standard deviation of X.

Average - ρ	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)					
IC models	Resnik	Lin	$J\&C$	$P\&S$	FaITH	Meng_{12}	Garla	$cosJ\&C$	Li_{s9}	Zhou	Meng_{14}	Gao	$coswJ\&C$	Avg
Seco et al. (2004)	0.6987	**0.7323**	0.7375	**0.7423**	**0.7323**	**0.7323**	0.7375	0.7375	0.7231	**0.7515**	0.7423	0.7294	0.7416	0.7337
Yuan et al. (2013)	0.6997	0.7284	0.7363	0.7402	0.7284	0.7284	0.7363	0.7363	0.7242	0.7430	0.7407	0.7294	0.7395	0.7316
Zhou et al. (2008a)	0.6974	0.7250	0.7353	0.7267	0.7250	0.7250	0.7353	0.7353	**0.7247**	0.7459	0.7393	0.7294	0.7488	0.7302
Meng et al. (2012)	0.6961	0.7204	0.7387	0.7317	0.7204	0.7204	0.7387	0.7387	0.7194	0.7417	0.7366	0.7294	0.7541	0.7297
Adhikari et al. (2015)	0.6960	0.7206	0.7362	0.7292	0.7206	0.7206	0.7362	0.7362	0.7196	0.7417	0.7359	0.7294	0.7521	0.7288
Sánchez et al. (2011)	0.7008	0.7266	**0.7402**	0.6804	0.7266	0.7266	**0.7402**	**0.7402**	0.7128	0.7411	0.7411	0.7317	**0.7579**	0.7282
CPRefCosine	0.6963	0.7250	0.7297	0.7131	0.7250	0.7250	0.7297	0.7297	0.7238	0.7344	0.7419	0.7285	0.7394	0.7263
Sánchez and Batet (2012)	0.6957	0.7264	0.7281	0.7148	0.7264	0.7264	0.7281	0.7281	0.7196	0.7312	0.7386	0.7274	0.7367	0.7252
Blanchard et al. (2008)	0.6973	0.7242	0.7274	0.7131	0.7242	0.7242	0.7274	0.7274	0.7182	0.7340	0.7387	0.7278	0.7365	0.7247
Resnik (1995)	0.7005	0.7219	0.7303	0.7008	0.7219	0.7219	0.7303	0.7303	0.7167	0.7322	**0.7437**	**0.7357**	0.7344	0.7247
CPRefHyponyms	0.7007	0.7222	0.7255	0.7112	0.7222	0.7222	0.7255	0.7255	0.7211	0.7285	0.7399	0.7323	0.7403	0.7244
CPCosine	0.6923	0.7229	0.7300	0.7124	0.7229	0.7229	0.7283	0.7300	0.7132	0.7345	0.7370	0.7313	0.7384	0.7243
CPRefLeaves	0.6990	0.7242	0.7247	0.7104	0.7242	0.7242	0.7247	0.7247	0.7200	0.7325	0.7394	0.7269	0.7411	0.7243
CPHyponyms	0.6897	0.7202	0.7350	0.7128	0.7202	0.7202	0.7333	0.7350	0.7140	0.7385	0.7346	0.7212	0.7403	0.7242
CPRefCosineLeaves	0.6960	0.7210	0.7261	0.7114	0.7210	0.7210	0.7261	0.7261	0.7235	0.7320	0.7400	0.7311	0.7375	0.7241
CPLeaves	0.6882	0.7198	0.7262	0.7080	0.7198	0.7198	0.7245	0.7262	0.7159	0.7313	0.7341	0.7243	0.7411	0.7215
CPRefLeaSubRatio	**0.7030**	0.7209	0.7169	0.7028	0.7209	0.7209	0.7169	0.7169	0.7200	0.7213	0.7323	0.7306	0.7342	0.7198
CPLogistic	0.6938	0.7178	0.7189	0.6888	0.7178	0.7178	0.7190	0.7189	0.7173	0.7190	0.7328	0.7242	0.7189	0.7158
CPRefCorpus	0.6940	0.7144	0.7105	0.6923	0.7144	0.7144	0.7105	0.7105	0.7224	0.7160	0.7386	0.7356	0.7262	0.7154
CPCorpus	0.6893	0.7076	0.7134	0.6858	0.7076	0.7076	0.7111	0.7134	0.7166	0.7181	0.7385	0.7348	0.7262	0.7131
CPRefLogisticLeaves	0.6711	0.6940	0.7150	0.6740	0.6940	0.6940	0.7150	0.7150	0.7162	0.7198	0.7228	0.7078	0.7353	0.7057
CPRefLogistic	0.6714	0.6936	0.7107	0.6713	0.6936	0.6936	0.7107	0.7107	0.7153	0.7193	0.7232	0.7090	0.7318	0.7042
CPRefUniform	0.6604	0.7103	0.6806	0.6481	0.7103	0.7103	0.6806	0.6806	0.7133	0.6846	0.7299	0.7186	0.7324	0.6969
CPUniform	0.6166	0.6987	0.6818	0.6063	0.6987	0.6987	0.6844	0.6817	0.7152	0.6845	0.7135	0.7170	0.7324	0.6869
Hadj Taieb et al. (2014a)	0.5364	0.6570	0.4627	0.5164	0.6570	0.6570	0.4627	0.4627	0.7030	0.4685	0.6964	0.7136	0.4627	0.5735
Best column value	0.7030	0.7323	0.7402	0.7423	0.7323	0.7323	0.7402	0.7402	0.7247	0.7515	0.7437	0.7357	**0.7579**	0.7337
Average per column*	0.6893	0.7183	0.7231	0.7012	0.7183	0.7183	0.7229	0.7231	0.7186	0.7282	0.7356	0.7268	0.7382	
Std. deviation*	0.0193	0.0101	0.0158	0.0289	0.0101	0.0101	0.0154	0.0158	0.0036	0.0160	0.0072	0.0077	0.0092	
Peak ratio*	0.7066	1.3971	1.0843	1.4244	1.3967	1.3971	1.1229	1.0841	1.7198	1.4520	1.1188	1.1668	2.1322	

Table 11: Average on all datasets of the Spearman (ρ) correlation values for each pair (IC model, IC measure). The IC models in bold are the new methods introduced in this work. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score on all datasets for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column. The baseline is defined by the corpus-based Resnik IC model. The rows are arranged in descending order according to the average score in last column. (*) The Hadj Taieb et al. (2014a) IC model can be considered as an outlier, thus, we excluded it from the computation of these statistics. The peak ratio (pr) for a random variable X is defined by the equation $pr(X) = \frac{|max(X)-\overline{X}|}{\sigma(X)}$, $\sigma(X)$ being the standard deviation of X.

		Pearson (r)					Spearman (ρ)					Overall average scores		
Sim. measures	IC model	RG65	MC28	Agirre201	P&S$_{full}$	SimLex	RG65	MC28	Agirre201	P&S$_{full}$	SimLex	r	ρ	both
coswJ&C (L&G)	Sánchez (2011)	**.8770**	**.8710**	.6933	.8850	.5918	**.8352**	**.8773**	**.6670**	**.8226**	.5873	.7836	**.7579**	**.7708**
Zhou et al	Seco et al	.8728	.8541	.6839	.8949	**.6237**	.8245	.8466	.6619	.8144	**.6101**	**.7859**	.7515	.7687
cosJ&C (L&G)	Sánchez (2011)	.8752	.8476	.6890	.8996	.5941	.8034	.8492	.6576	.8003	.5906	.7811	.7402	.7607
Pirró & Seco	Seco et al	.8622	.8463	.6890	.8970	.5975	.8012	.8678	.6643	.7919	.5862	.7784	.7423	.7603
Taieb et al	not apply	.8670	.8248	**.7123**	**.9068**	.6093	.7972	.8077	.6633	.7973	.5960	.7840	.7323	.7582
Gao et al	Resnik	.8709	.8336	.6731	.8909	.6152	.8153	.8082	.6469	.8089	.5994	.7768	.7357	.7563
Jiang & Conrath	Sánchez (2011)	.8619	.8595	.6591	.8762	.5838	.8034	.8492	.6576	.8003	.5906	.7681	.7402	.7542
Meng & Gu	Seco et al	.8596	.8144	.6969	.9031	.6048	.7972	.8314	.6530	.7911	.5888	.7758	.7323	.7540
P&S FaITH	Seco et al	.8565	.8094	.6966	.9042	.6046	.7972	.8314	.6530	.7911	.5888	.7743	.7323	.7533
Lin	Seco et al	.8609	.8240	.6850	.8945	.6010	.7972	.8314	.6530	.7911	.5888	.7731	.7323	.7527
Li$_{s3}$ et al	not apply	.8625	.8355	.6675	.8853	.6059	.8106	.8144	.6313	.7986	.5918	.7713	.7294	.7504
Li$_{s9}$ et al	Zhou et al	.8438	.8102	.6999	.8897	.5912	.7932	.7986	.6612	.7903	.5804	.7669	.7247	.7458
Li$_{s4}$ et al	not apply	.8598	.8281	.6669	.8787	.6052	.7970	.7729	.6443	.7879	.5875	.7677	.7179	.7428
Sánchez et al	not apply	.8477	.8062	.6751	.8703	.5940	.7843	.7908	.6505	.7895	.5785	.7587	.7187	.7387
Meng et al	Resnik	.8434	.8064	.6006	.8286	.5659	.8227	.8135	.6663	.8105	.6057	.7290	.7437	.7363
Al-Mubaid	not apply	.8075	.7906	.6503	.8530	.5756	.8123	.8056	.6515	.8070	.5778	.7354	.7308	.7331
Resnik	**CPRefLSRat**	.8232	.7934	.6727	.8740	.5124	.7641	.8395	.6424	.7578	.5112	.7352	.7030	.7191
Pedersen et al	not apply	.7807	.7585	.6064	.8398	.5509	.8106	.8144	.6313	.7986	.5918	.7072	.7294	.7183
Leacock & Ch.	not apply	.7939	.7538	.5950	.7777	.5740	.8106	.8144	.6313	.7986	.5918	.6989	.7294	.7141
Garla & Brandt	Sánchez (2011)	.7690	.7198	.5733	.8470	.5117	.8034	.8492	.6576	.8003	.5906	.6842	.7402	.7122
Rada et al	not apply	.7708	.7294	.5798	.7506	.5651	.8106	.8144	.6313	.7986	.5918	.6791	.7294	.7043
Wu & Palmer	not apply	.7703	.7746	.6048	.7761	.5374	.7492	.7525	.6368	.7458	.5417	.6926	.6852	.6889
Best column values		.8770	.8710	.7123	.9068	.6237	.8352	.8773	.6670	.8226	.6101	.7859	.7579	.7708

Table 12: Results for each similarity measure on each dataset: Pearson (r) and Spearman (ρ) correlation coefficients, and averaged overall scores. Each IC-based similarity measures is evaluated with its best performing IC model in average, in accordance with the average Spearman correlation values in table 9. The bold values represent the best score within each column. The rows are arranged in descending order according to the overall average score in last column. The baseline is defined by the Jiang-Conrath similarity measure.

IC models	Resnik (1995)	Seco et al. (2004)	Blanchard et al. (2008)	Zhou et al. (2008a)	Sánchez et al. (2011)	Sánchez and Batet (2012)	Meng et al. (2012)	Yuan et al. (2013)	Hadj Taieb et al. (2014a)	Adhikari et al. (2015)	CondProbHyponyms	CondProbUniform	CondProbLeaves	CondProbCosine	CondProbLogistic	CondProbCorpus	**CondProbRefHyponyms**	**CondProbRefUniform**	**CondProbRefLeaves**	**CondProbRefCosine**	**CondProbRefLogistic**	**CondProbRefCorpus**	**CondProbRefCosineLeaves**	**CondProbRefLogisticLeaves**	**CPRefLeavesSubsumersRatio**
Resnik (1995)	—								x			x	x		x	x		x			x	x		x	x
Seco et al. (2004)	x	—	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Blanchard et al. (2008)			—						x			x	x		x	x		x		x	x	x		x	x
Zhou et al. (2008a)	x		x	—		x			x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Sánchez et al. (2011)					—				x			x	x		x	x		x			x	x		x	x
Sánchez and Batet (2012)						—			x			x	x		x	x		x			x	x		x	x
Meng et al. (2012)			x			x	—		x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x
Yuan et al. (2013)	x		x			x		—	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Hadj Taieb et al. (2014a)									—																
Adhikari et al. (2015)			x						x	—	x	x	x	x	x	x	x	x	x		x	x	x	x	x
CondProbHyponyms									x		—	x	x		x	x		x			x	x		x	
CondProbUniform									x			—													
CondProbLeaves									x			x	—		x	x		x			x	x		x	
CondProbCosine									x			x	x	—	x	x		x			x	x		x	x
CondProbLogistic									x			x			—			x			x			x	
CondProbCorpus									x			x				—		x			x			x	
CondProbRefHyponyms									x			x	x		x	x	—	x			x	x		x	x
CondProbRefUniform									x			**x**						—							
CondProbRefLeaves									x			x	**x**		x	x		x	—		x	x		x	x
CondProbRefCosine			x			x			x			x	x	**x**	x	x	x	x	x	—	x	x	x	x	x
CondProbRefLogistic									x			x									—				
CondProbRefCorpus									x			x				**x**		x			x	—		x	
CondProbRefCosineLeaves									x			x	x		x	x		x			x	x	—	x	x
CondProbRefLogisticLeaves									x			x									x			—	
CPRefLeavesSubsumersRatio									x			x			x	x		x			x	x		x	—

Table 13: Summary of the statistical significance analysis between IC models derived from the pairwise raw p-values shown in table 25. The p-values are computed using a one-sided t-student distribution for the paired vectors of Spearman correlation values reported by each pair of IC models on each dataset. Each row shows a 'x' whenever the row IC model obtains a statistically significant higher performance than the column IC model. Thus, the rows show the IC models that are outperformed by each IC model on the left, whilst the columns show the IC models that ourperform each IC model on the top. For instance, the Seco et al. (2004) IC model outperforms all IC models with the only exception of the Sánchez et al. (2011) IC model. On the other hand, a first glance to the columns shows that the Seco et al. (2004) and Sánchez et al. (2011) IC models are the only ones that are not outperformed by another IC model. Bold values in uppercase 'X' show the outperformance of the refined IC models as regard their corresponding non-refined models proving the main hypothesis.

Similarity measures	Lastra & García (2015), coswJ&C	Zhou et al. (2008b)	Lastra & García (2015), cosJ&C	Pirró and Seco (2008)	Hadj Taieb et al. (2014b)	Gao et al. (2015)	Jiang and Conrath (1997)	Meng and Gu (2012)	Pirró and Euzenat (2010)	Lin (1998)	Li et al. (2003), strat. 3	Li et al. (2003), strat. 9	Li et al. (2003), strat. 4	Sánchez et al. (2012)	Meng et al. (2014)	Al-Mubaid and Nguyen (2009)	Resnik (1995)	Pedersen et al. (2007)	Leacock and Chodorow (1998)	Garla and Brandt (2012)	Rada et al. (1989)	Wu and Palmer (1994)
Lastra & García (2015), coswJ&C	—		x	x			x	x	x	x	x	x	x	x		x	x	x	x	x	x	x
Zhou et al. (2008b)		—	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x
Lastra & García (2015), cosJ&C			—					x	x	x				x			x					x
Pirró and Seco (2008)				—													x					x
Hadj Taieb et al. (2014b)					—							x	x	x								x
Gao et al. (2015)						—							x	x								x
Jiang and Conrath (1997)							—	x	x	x				x			x					x
Meng and Gu (2012)								—														x
Pirró and Euzenat (2010)									—													x
Lin (1998)										—												x
Li et al. (2003), strat. 3											—											x
Li et al. (2003), strat. 9												—		x								x
Li et al. (2003), strat. 4													—									x
Sánchez et al. (2012)														—								x
Meng et al. (2014)					x	x					x	x	x	x	—	x		x	x		x	x
Al-Mubaid and Nguyen (2009)														x		—						x
Resnik (1995)																	—					
Pedersen et al. (2007)																		—				x
Leacock and Chodorow (1998)																			—			x
Garla and Brandt (2012)								x	x	x				x			x			—		x
Rada et al. (1989)																					—	x
Wu and Palmer (1994)																						—

Table 14: Summary of the statistical significance analysis between similarity measures derived from the pairwise raw p-values shown in table 26. The p-values are computed using a one-sided t-student distribution for the paired vectors of Spearman correlation values reported by each pair of similarity measures on each dataset. Each row shows a 'x' whenever the row measure obtains a statistically significant higher performance than the respective column measure. Thus, the rows show the similarity measures that are outperformed by each measure on the left, whilst the columns show the similarity measures that ourperform the measures on the top. For instance, the coswJ&C and Zhou et al. (2008b) similarity measures outperform most similarity measures, whlist a first glance to the columns show that the Wu and Palmer (1994) measure is outperformed by the remaining similarity measures.

RG65 - r	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
IC models	Resnik	Lin	*J&C*	*P&S*	FaITH	Meng_{12}	Garla	*cosJ&C*	Li_{s9}	Zhou	Meng_{14}	Gao	*coswJ&C*
Resnik (1995)	0.8345	0.8589	0.8561	0.8335	0.8585	0.8609	0.6950	0.8653	0.8617	0.8582	0.8434	**0.8709**	0.8703
Seco et al. (2004)	0.8326	0.8609	0.8546	**0.8622**	0.8565	0.8596	0.8571	0.8642	0.8241	**0.8728**	**0.8486**	0.7992	0.8634
Blanchard et al. (2008)	0.8310	0.8589	0.8493	0.8303	0.8536	0.8571	0.6957	0.8607	0.8618	0.8525	0.8482	0.8669	0.8614
Zhou et al. (2008a)	0.8080	0.8259	0.8286	0.8334	0.8589	0.8539	0.8662	0.8558	0.8438	0.8574	0.7749	0.7992	0.8610
Sánchez et al. (2011)	**0.8409**	0.8530	**0.8619**	0.8105	**0.8683**	0.8663	0.7690	**0.8752**	0.8586	0.8639	0.8147	0.8682	**0.8770**
Sánchez and Batet (2012)	0.8355	0.8616	0.8508	0.8332	0.8565	0.8600	0.6940	0.8606	0.8615	0.8535	0.8475	0.8678	0.8614
Meng et al. (2012)	0.8260	0.8608	0.8598	0.8586	0.8658	0.8670	**0.8717**	0.8723	0.8282	0.8638	0.8285	0.7992	0.8747
Yuan et al. (2013)	0.8243	0.8621	0.8505	0.8607	0.8649	0.8675	0.8609	0.8632	0.8231	0.8629	0.8273	0.7992	0.8624
Hadj Taieb et al. (2014a)	0.4658	0.7825	0.4543	0.5090	0.7933	0.7924	0.5939	0.4552	0.8444	0.4586	0.7693	0.8527	0.4552
Adhikari et al. (2015)	0.8264	0.8612	0.8609	0.8592	0.8650	0.8664	0.8707	0.8722	0.8267	0.8650	0.8321	0.7992	0.8747
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.8283	0.8587	0.8562	0.8286	0.8556	0.8585	0.7033	0.8658	0.8606	0.8586	0.8426	0.8665	0.8645
CPUniform	0.6844	0.7958	0.7228	0.6840	0.8425	0.8323	0.6721	0.7473	0.8549	0.7277	0.7520	0.8569	0.8331
CPLeaves	0.8272	0.8575	0.8535	0.8263	0.8550	0.8578	0.7009	0.8635	0.8605	0.8560	0.8418	0.8662	0.8625
CPCosine	0.8294	0.8591	0.8540	0.8265	0.8550	0.8581	0.6609	0.8634	0.8574	0.8555	0.8456	0.8658	0.8624
CPLogistic	0.8021	**0.8681**	0.8260	0.8147	0.8652	**0.8692**	0.6899	0.8501	0.8605	0.8300	0.8363	0.8656	0.8595
CPCorpus	0.8296	0.8560	0.8527	0.8223	0.8575	0.8591	0.7275	0.8633	0.8618	0.8552	0.8405	0.8706	0.8682
IC models introduced in this work													
CPRefHyponyms	0.8242	0.8547	0.8497	0.8256	0.8514	0.8543	0.7001	0.8610	0.8615	0.8527	0.8463	0.8648	0.8645
CPRefUniform	0.7253	0.8540	0.7484	0.7453	0.8651	0.8652	0.6810	0.7703	0.8551	0.7530	0.8103	0.8615	0.8331
CPRefLeaves	0.8230	0.8533	0.8469	0.8231	0.8508	0.8535	0.6980	0.8587	0.8614	0.8499	0.8454	0.8644	0.8625
CPRefCosine	0.8231	0.8573	0.8506	0.8252	0.8532	0.8563	0.7115	0.8609	0.8617	0.8539	0.8467	0.8663	0.8621
CPRefLogistic	0.7991	0.8440	0.8421	0.8125	0.8536	0.8535	0.7173	0.8597	0.8502	0.8454	0.8355	0.8444	0.8648
CPRefCorpus	0.8264	0.8528	0.8484	0.8210	0.8532	0.8550	0.7217	0.8588	**0.8627**	0.8512	0.8448	0.8684	0.8682
CPRefCosineLeaves	0.8221	0.8546	0.8457	0.8210	0.8517	0.8544	0.7044	0.8568	0.8613	0.8491	0.8451	0.8656	0.8587
CPRefLogisticLeaves	0.7991	0.8430	0.8427	0.8120	0.8535	0.8532	0.7135	0.8599	0.8495	0.8458	0.8336	0.8438	0.8653
CPRefLeavesSubsumersRatio	0.8232	0.8564	0.8543	0.8292	0.8535	0.8564	0.6850	0.8621	0.8614	0.8560	0.8459	0.8649	0.8682
Best column value	0.8409	0.8681	0.8619	0.8622	0.8683	0.8692	0.8717	0.8752	0.8627	0.8728	0.8486	0.8709	**0.8770**

Table 15: Pearson (r) correlation coefficients for all the IC models and measures in the RG65 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

RG65 - ρ	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
IC models	Resnik	Lin	*J&C*	*P&S*	FaITH	Meng_{12}	Garla	*cosJ&C*	Li_{s9}	Zhou	Meng_{14}	Gao	*coswJ&C*
Resnik (1995)	0.7777	0.7761	0.7831	0.7633	0.7761	0.7761	0.7831	0.7831	0.7868	0.7848	**0.8227**	**0.8153**	0.8009
Seco et al. (2004)	0.7735	**0.7972**	0.7866	0.8012	**0.7972**	**0.7972**	0.7866	0.7866	**0.7939**	**0.8245**	0.8204	0.8106	0.8061
Blanchard et al. (2008)	0.7602	0.7850	0.7788	0.7816	0.7850	0.7850	0.7788	0.7788	0.7867	0.7855	0.8174	0.7980	0.8014
Zhou et al. (2008a)	0.7690	0.7881	0.8051	0.7958	0.7881	0.7881	0.8051	0.8051	0.7932	0.8244	0.8147	0.8106	0.8244
Sánchez et al. (2011)	0.7714	0.7944	0.8034	0.7671	0.7944	0.7944	0.8034	0.8034	0.7846	0.8081	0.8201	0.8108	0.8352
Sánchez and Batet (2012)	0.7706	0.7911	0.7779	0.7856	0.7911	0.7911	0.7779	0.7779	0.7869	0.7854	0.8170	0.8000	0.8003
Meng et al. (2012)	0.7607	0.7817	**0.8166**	0.8140	0.7817	0.7817	**0.8166**	**0.8166**	0.7886	0.8177	0.8131	0.8106	**0.8408**
Yuan et al. (2013)	0.7742	0.7919	0.8050	**0.8206**	0.7919	0.7919	0.8050	0.8050	0.7932	0.8195	0.8151	0.8106	0.8174
Hadj Taieb et al. (2014a)	0.5990	0.7417	0.6126	0.6123	0.7417	0.7417	0.6126	0.6126	0.7719	0.6182	0.7817	0.7787	0.6126
Adhikari et al. (2015)	0.7609	0.7829	0.8086	0.8084	0.7829	0.7829	0.8086	0.8086	0.7886	0.8178	0.8114	0.8106	0.8369
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.7561	0.7888	0.8017	0.7941	0.7888	0.7888	0.8017	0.8017	0.7826	0.8052	0.8154	0.7959	0.8071
CPUniform	0.7003	0.7786	0.7613	0.6930	0.7786	0.7786	0.7613	0.7613	0.7888	0.7638	0.7928	0.7880	0.8146
CPLeaves	0.7549	0.7868	0.7877	0.7848	0.7868	0.7868	0.7877	0.7877	0.7825	0.7951	0.8143	0.7965	0.8081
CPCosine	0.7710	0.7896	0.7835	0.7830	0.7896	0.7896	0.7835	0.7835	0.7874	0.7882	0.8174	0.8116	0.8037
CPLogistic	**0.7833**	0.7966	0.7993	0.7792	0.7966	0.7966	0.7993	0.7993	0.7899	0.8003	0.8145	0.7984	0.7977
CPCorpus	0.7702	0.7652	0.7722	0.7515	0.7652	0.7652	0.7722	0.7722	0.7868	0.7786	0.8208	0.8148	0.7833
IC models introduced in this work													
CPRefHyponyms	0.7651	0.7786	0.7765	0.7791	0.7786	0.7786	0.7765	0.7765	0.7882	0.7812	0.8142	0.8018	0.8071
CPRefUniform	0.7273	0.7773	0.7475	0.7220	0.7773	0.7773	0.7475	0.7475	0.7826	0.7526	0.8039	0.7841	0.8146
CPRefLeaves	0.7627	0.7804	0.7760	0.7779	0.7804	0.7804	0.7760	0.7760	0.7875	0.7861	0.8125	0.7880	0.8081
CPRefCosine	0.7606	0.7893	0.7835	0.7794	0.7893	0.7893	0.7835	0.7835	0.7911	0.7878	0.8197	0.7996	0.8137
CPRefLogistic	0.7348	0.7404	0.7645	0.7320	0.7404	0.7404	0.7645	0.7645	0.7570	0.7756	0.7782	0.7492	0.8044
CPRefCorpus	0.7612	0.7635	0.7580	0.7482	0.7635	0.7635	0.7580	0.7580	0.7897	0.7649	0.8131	0.8021	0.7833
CPRefCosineLeaves	0.7600	0.7796	0.7752	0.7753	0.7796	0.7796	0.7752	0.7752	0.7920	0.7825	0.8158	0.8013	0.8061
CPRefLogisticLeaves	0.7347	0.7407	0.7732	0.7351	0.7407	0.7407	0.7732	0.7732	0.7587	0.7785	0.7772	0.7485	0.8112
CPRefLeavesSubsumersRatio	0.7641	0.7728	0.7810	0.7724	0.7728	0.7728	0.7810	0.7810	0.7877	0.7867	0.8090	0.7965	0.8129
Best column value	0.7833	0.7972	0.8166	0.8206	0.7972	0.7972	0.8166	0.8166	0.7939	0.8245	0.8227	0.8153	**0.8408**

Table 16: Spearman (ρ) correlation coefficients for all the IC models and measures in the RG65 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

MC28 - r	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
IC models	Resnik	Lin	*J&C*	*P&S*	FaITH	Meng_{12}	Garla	*cosJ&C*	Li_{s9}	Zhou	Meng_{14}	Gao	*coswJ&C*
Resnik (1995)	0.7929	0.8350	**0.8809**	**0.8466**	0.8233	0.8276	0.6611	**0.8710**	0.8260	**0.8795**	0.8064	0.8336	**0.8789**
Seco et al. (2004)	0.7834	0.8240	0.8557	0.8463	0.8094	0.8144	0.8299	0.8427	0.7730	0.8541	0.8107	0.7775	0.8369
Blanchard et al. (2008)	0.7823	0.8228	0.8528	0.8169	0.8060	0.8118	0.6694	0.8385	0.8267	0.8535	0.8122	0.8299	0.8356
Zhou et al. (2008a)	0.8179	0.8222	0.8282	0.8281	**0.8344**	0.8338	**0.8410**	0.8403	0.8102	0.8396	0.7638	0.7775	0.8492
Sánchez et al. (2011)	0.8189	0.8357	0.8595	0.8056	0.8343	0.8359	0.7198	0.8476	0.8236	0.8582	0.7912	0.8366	0.8710
Sánchez and Batet (2012)	0.7905	0.8261	0.8507	0.8177	0.8103	0.8159	0.6686	0.8411	0.8247	0.8511	0.8101	0.8282	0.8384
Meng et al. (2012)	0.8202	0.8393	0.8314	0.8330	0.8330	0.8361	0.8376	0.8393	0.7959	0.8350	0.8005	0.7775	0.8569
Yuan et al. (2013)	0.8084	0.8341	0.8347	0.8384	0.8277	0.8315	0.8320	0.8407	0.7881	0.8381	0.7953	0.7775	0.8350
Hadj Taieb et al. (2014a)	0.5391	0.6842	0.4587	0.4845	0.6899	0.6875	0.5516	0.4594	0.8049	0.4622	0.6792	0.8243	0.4594
Adhikari et al. (2015)	**0.8211**	**0.8410**	0.8331	0.8351	0.8331	**0.8365**	0.8366	0.8392	0.7949	0.8368	0.8040	0.7775	0.8552
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.7860	0.8215	0.8552	0.8110	0.8070	0.8124	0.6687	0.8429	0.8242	0.8548	0.8030	0.8280	0.8406
CPUniform	0.7408	0.7753	0.6483	0.6365	0.8039	0.7971	0.6250	0.6698	0.8190	0.6530	0.7276	0.8325	0.7925
CPLeaves	0.7869	0.8219	0.8511	0.8073	0.8080	0.8131	0.6694	0.8402	0.8248	0.8507	0.8039	0.8287	0.8387
CPCosine	0.7797	0.8208	0.8562	0.8127	0.8054	0.8109	0.6342	0.8445	0.8211	0.8563	0.8058	0.8316	0.8410
CPLogistic	0.7700	0.8142	0.8020	0.7687	0.8079	0.8119	0.6624	0.8095	0.8225	0.8046	0.7801	0.8249	0.8232
CPCorpus	0.7912	0.8290	0.8678	0.8281	0.8163	0.8209	0.6839	0.8613	0.8262	0.8668	0.8014	0.8338	0.8745
IC models introduced in this work													
CPRefHyponyms	0.7884	0.8242	0.8519	0.8134	0.8061	0.8122	0.6670	0.8412	0.8315	0.8522	0.8170	0.8348	0.8406
CPRefUniform	0.7317	0.8257	0.6998	0.7000	0.8223	0.8254	0.6487	0.7206	0.8193	0.7035	0.7782	0.8344	0.7925
CPRefLeaves	0.7891	0.8244	0.8476	0.8092	0.8073	0.8131	0.6675	0.8386	0.8322	0.8480	0.8177	0.8355	0.8387
CPRefCosine	0.7767	0.8215	0.8533	0.8144	0.8029	0.8093	0.6675	0.8455	0.8278	0.8540	0.8123	0.8309	0.8393
CPRefLogistic	0.7848	0.8226	0.8376	0.7999	0.8183	0.8210	0.6818	0.8374	**0.8374**	0.8396	0.8248	0.8288	0.8437
CPRefCorpus	0.7950	0.8312	0.8675	0.8324	0.8155	0.8206	0.6758	0.8597	0.8325	0.8670	0.8165	**0.8383**	0.8745
CPRefCosineLeaves	0.7805	0.8220	0.8456	0.8069	0.8046	0.8106	0.6680	0.8391	0.8288	0.8464	0.8141	0.8326	0.8365
CPRefLogisticLeaves	0.7881	0.8240	0.8383	0.8004	0.8202	0.8227	0.6836	0.8385	0.8374	0.8402	**0.8252**	0.8292	0.8444
CPRefLeavesSubsumersRatio	0.7934	0.8243	0.8363	0.7975	0.8115	0.8160	0.6565	0.8356	0.8315	0.8362	0.8143	0.8352	0.8417
Best column value	0.8211	0.8410	**0.8809**	0.8466	0.8344	0.8365	0.8410	0.8710	0.8374	0.8795	0.8252	0.8383	0.8789

Table 17: Pearson (r) correlation coefficients for all the IC models and measures in the MC28 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

MC28 - ρ	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
IC models	Resnik	Lin	*J&C*	*P&S*	FaITH	Meng_{12}	Garla	*cosJ&C*	Li_{s9}	Zhou	Meng_{14}	Gao	*coswJ&C*
Resnik (1995)	0.7935	**0.8403**	**0.8882**	0.8255	**0.8403**	**0.8403**	**0.8882**	**0.8882**	0.7734	**0.8871**	**0.8135**	0.8082	**0.8923**
Seco et al. (2004)	0.7947	0.8314	0.8727	**0.8678**	0.8314	0.8314	0.8727	0.8727	0.7956	0.8466	0.8053	0.8144	0.8801
Blanchard et al. (2008)	0.8239	0.8225	0.8518	0.8335	0.8225	0.8225	0.8518	0.8518	0.7824	0.8643	0.7973	0.8112	0.8749
Zhou et al. (2008a)	0.7971	0.8072	0.8244	0.8091	0.8072	0.8072	0.8244	0.8244	0.7986	0.8192	0.8042	0.8144	0.8600
Sánchez et al. (2011)	0.7937	0.8114	0.8492	0.7348	0.8114	0.8114	0.8492	0.8492	0.7635	0.8413	0.8058	0.7983	0.8773
Sánchez and Batet (2012)	0.7774	0.8212	0.8551	0.8329	0.8212	0.8212	0.8551	0.8551	0.7865	0.8518	0.7968	0.7972	0.8734
Meng et al. (2012)	0.8296	0.8080	0.8198	0.8050	0.8080	0.8080	0.8198	0.8198	0.7926	0.8198	0.8042	0.8144	0.8543
Yuan et al. (2013)	0.7971	0.8042	0.8274	0.8083	0.8042	0.8042	0.8274	0.8274	0.7975	0.8209	0.8031	0.8144	0.8285
Hadj Taieb et al. (2014a)	0.6340	0.6961	0.6102	0.4673	0.6961	0.6961	0.6102	0.6102	0.7882	0.6110	0.7429	0.8010	0.6102
Adhikari et al. (2015)	0.8296	0.8080	0.8198	0.8050	0.8080	0.8080	0.8198	0.8198	0.7926	0.8198	0.8036	0.8144	0.8526
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.8131	0.8034	0.8554	0.8187	0.8034	0.8034	0.8554	0.8554	0.7734	0.8589	0.7875	0.7842	0.8753
CPUniform	0.7564	0.7749	0.7281	0.6155	0.7749	0.7749	0.7281	0.7281	0.7761	0.7339	0.7746	0.7813	0.8077
CPLeaves	0.8073	0.8039	0.8389	0.8110	0.8039	0.8039	0.8389	0.8389	0.7824	0.8436	0.7875	0.7979	0.8741
CPCosine	0.7791	0.8116	0.8606	0.8324	0.8116	0.8116	0.8606	0.8606	0.7635	0.8688	0.7919	0.8012	0.8751
CPLogistic	0.7909	0.7752	0.8034	0.7377	0.7752	0.7752	0.8034	0.8034	0.7770	0.7968	0.7848	0.7957	0.8097
CPCorpus	0.7707	0.8058	0.8502	0.7946	0.8058	0.8058	0.8502	0.8502	0.7734	0.8537	0.7984	0.8075	0.8797
IC models introduced in this work													
CPRefHyponyms	**0.8433**	0.8313	0.8504	0.8274	0.8313	0.8313	0.8504	0.8504	0.7958	0.8507	0.8116	0.8276	0.8753
CPRefUniform	0.7738	0.7713	0.7287	0.6597	0.7713	0.7713	0.7287	0.7287	0.7729	0.7344	0.7837	0.7868	0.8077
CPRefLeaves	0.8376	0.8362	0.8409	0.8242	0.8362	0.8362	0.8409	0.8409	0.7912	0.8573	0.8116	0.8276	0.8741
CPRefCosine	0.8288	0.8217	0.8510	0.8296	0.8217	0.8217	0.8510	0.8510	0.8049	0.8600	0.8067	0.8134	0.8693
CPRefLogistic	0.7737	0.7757	0.8124	0.7547	0.7757	0.7757	0.8124	0.8124	0.8253	0.8280	0.8108	0.8135	0.8461
CPRefCorpus	0.8124	0.8368	0.8592	0.8181	0.8368	0.8368	0.8592	0.8592	0.7958	0.8649	0.8110	**0.8333**	0.8797
CPRefCosineLeaves	0.8310	0.8217	0.8466	0.8296	0.8217	0.8217	0.8466	0.8466	0.7999	0.8567	0.8077	0.8221	0.8655
CPRefLogisticLeaves	0.7737	0.7757	0.8138	0.7596	0.7757	0.7757	0.8138	0.8138	**0.8258**	0.8239	0.8119	0.8086	0.8452
CPRefLeavesSubsumersRatio	0.8395	0.8395	0.7960	0.7968	0.8395	0.8395	0.7960	0.7960	0.7895	0.8039	0.7905	0.8248	0.8354
Best column value	0.8433	0.8403	0.8882	0.8678	0.8403	0.8403	0.8882	0.8882	0.8258	0.8871	0.8135	0.8333	**0.8923**

Table 18: Spearman (ρ) correlation coefficients for all the IC models and measures in the MC28 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

Agirre201 - r	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
IC models	Resnik	Lin	*J&C*	*P&S*	FaITH	Meng_{12}	Garla	*cosJ&C*	Li_{s9}	Zhou	Meng_{14}	Gao	*coswJ&C*
Resnik (1995)	0.6716	0.6697	0.6349	0.6401	0.6913	0.6882	0.5007	0.6524	0.6726	0.6393	0.6006	0.6731	0.6490
Seco et al. (2004)	0.6629	0.6850	**0.6724**	**0.6890**	0.6966	0.6969	0.6885	**0.6904**	0.6856	**0.6839**	0.6185	0.6252	0.6784
Blanchard et al. (2008)	0.6615	0.6834	0.6716	0.6590	0.6939	0.6945	0.5046	0.6891	0.6776	0.6742	0.6184	0.6751	0.6803
Zhou et al. (2008a)	0.6453	0.6409	0.6288	0.6503	0.6848	0.6753	0.6815	0.6564	**0.6999**	0.6616	0.5591	0.6252	0.6638
Sánchez et al. (2011)	0.6592	0.6643	0.6591	0.6339	0.6946	0.6889	0.5733	0.6890	0.6676	0.6616	0.5842	0.6664	**0.6933**
Sánchez and Batet (2012)	0.6678	0.6848	0.6720	0.6602	0.6972	0.6973	0.5052	0.6878	0.6762	0.6743	0.6167	0.6747	0.6788
Meng et al. (2012)	0.6756	0.6817	0.6593	0.6863	0.7039	0.7008	**0.6890**	0.6747	0.6980	0.6719	0.6006	0.6252	0.6782
Yuan et al. (2013)	**0.6760**	0.6858	0.6543	0.6863	0.7061	0.7040	0.6834	0.6695	0.6944	0.6681	0.6046	0.6252	0.6633
Hadj Taieb et al. (2014a)	0.4165	0.6345	0.0790	0.3150	0.6490	0.6467	0.2674	0.0793	0.6827	0.0844	0.5885	0.6774	0.0793
Adhikari et al. (2015)	0.6755	0.6824	0.6606	0.6866	0.7041	0.7013	0.6890	0.6756	0.6975	0.6728	0.6028	0.6252	0.6786
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.6557	0.6751	0.6585	0.6470	0.6868	0.6874	0.5003	0.6748	0.6753	0.6618	0.6097	0.6723	0.6752
CPUniform	0.5713	0.6156	0.5597	0.5682	0.6516	0.6432	0.4932	0.5719	0.6617	0.5636	0.5505	0.6535	0.6233
CPLeaves	0.6569	0.6765	0.6601	0.6478	0.6888	0.6891	0.1151	0.6764	0.6755	0.6631	0.6096	0.6725	0.6773
CPCosine	0.6579	0.6789	0.6660	0.6521	0.6893	0.6902	0.4471	0.6805	0.6660	0.6676	0.6144	0.6598	0.6735
CPLogistic	0.6501	0.6618	0.6206	0.6290	0.6809	0.6796	0.4783	0.6380	0.6716	0.6242	0.5959	0.6653	0.6416
CPCorpus	0.6638	0.6603	0.6285	0.6326	0.6807	0.6779	0.5162	0.6426	0.6727	0.6339	0.5939	0.6728	0.6526
IC models introduced in this work													
CPRefHyponyms	0.6604	0.6817	0.6646	0.6557	0.6915	0.6924	0.5025	0.6798	0.6787	0.6675	0.6178	0.6763	0.6752
CPRefUniform	0.6533	**0.6926**	0.5964	0.6423	**0.7137**	**0.7110**	0.4780	0.6086	0.6621	0.6002	0.6027	0.6618	0.6233
CPRefLeaves	0.6612	0.6827	0.6658	0.6563	0.6931	0.6938	0.5040	0.6813	0.6790	0.6686	0.6177	0.6764	0.6773
CPRefCosine	0.6586	0.6798	0.6614	0.6550	0.6884	0.6898	0.4998	0.6736	0.6772	0.6647	0.6172	0.6763	0.6591
CPRefLogistic	0.6558	0.6735	0.6543	0.6488	0.6926	0.6907	0.4904	0.6695	0.6857	0.6570	0.6148	**0.6794**	0.6641
CPRefCorpus	0.6707	0.6682	0.6348	0.6416	0.6864	0.6843	0.5150	0.6484	0.6775	0.6398	0.6024	0.6772	0.6526
CPRefCosineLeaves	0.6608	0.6819	0.6638	0.6564	0.6917	0.6927	0.5036	0.6771	0.6776	0.6670	0.6169	0.6766	0.6637
CPRefLogisticLeaves	0.6565	0.6742	0.6535	0.6488	0.6943	0.6921	0.4920	0.6688	0.6858	0.6561	0.6140	0.6793	0.6643
CPRefLeavesSubsumersRatio	0.6727	0.6877	0.6664	0.6603	0.6977	0.6984	0.4949	0.6757	0.6786	0.6683	**0.6191**	0.6767	0.6709
Best column value	0.6760	0.6926	0.6724	0.6890	**0.7137**	0.7110	0.6890	0.6904	0.6999	0.6839	0.6191	0.6794	0.6933

Table 19: Pearson (r) correlation coefficients for all the IC models and measures in the Agirre201 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

Agirre201 - ρ	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
IC models	Resnik	Lin	$J\&C$	$P\&S$	FaITH	Meng_{12}	Garla	$cosJ\&C$	Li_{s9}	Zhou	Meng_{14}	Gao	$coswJ\&C$
Resnik (1995)	0.6320	0.6461	0.6382	0.6282	0.6461	0.6461	0.6382	0.6382	0.6503	0.6412	0.6663	0.6469	0.6331
Seco et al. (2004)	0.6308	0.6530	**0.6612**	0.6643	0.6530	0.6530	**0.6612**	**0.6612**	0.6560	**0.6619**	0.6606	0.6313	0.6482
Blanchard et al. (2008)	0.6309	0.6505	0.6577	0.6409	0.6505	0.6505	0.6577	0.6577	0.6536	0.6613	0.6594	0.6445	0.6465
Zhou et al. (2008a)	0.6460	0.6591	0.6524	0.6581	0.6591	0.6591	0.6524	0.6524	0.6612	0.6556	0.6631	0.6313	0.6591
Sánchez et al. (2011)	0.6408	0.6534	0.6576	0.6224	0.6534	0.6534	0.6576	0.6576	0.6482	0.6598	0.6601	0.6496	**0.6670**
Sánchez and Batet (2012)	0.6345	0.6494	0.6590	0.6399	0.6494	0.6494	0.6590	0.6590	0.6545	0.6590	0.6606	0.6476	0.6490
Meng et al. (2012)	0.6462	0.6581	0.6488	0.6560	0.6581	0.6581	0.6488	0.6488	0.6646	0.6532	0.6620	0.6313	0.6562
Yuan et al. (2013)	**0.6481**	0.6652	0.6505	**0.6656**	0.6652	0.6652	0.6505	0.6505	**0.6656**	0.6495	**0.6678**	0.6313	0.6466
Hadj Taieb et al. (2014a)	0.5343	0.6175	0.1556	0.4676	0.6175	0.6175	0.1556	0.1556	0.6600	0.1672	0.6310	**0.6662**	0.1556
Adhikari et al. (2015)	0.6457	0.6584	0.6497	0.6563	0.6584	0.6584	0.6497	0.6497	0.6651	0.6538	0.6610	0.6313	0.6560
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.6267	0.6466	0.6462	0.6329	0.6466	0.6466	0.6460	0.6462	0.6521	0.6491	0.6553	0.6455	0.6452
CPUniform	0.5617	0.6325	0.6005	0.5704	0.6325	0.6325	0.5973	0.6002	0.6482	0.6022	0.6455	0.6478	0.6395
CPLeaves	0.6289	0.6478	0.6476	0.6363	0.6478	0.6478	0.6476	0.6476	0.6534	0.6509	0.6553	0.6462	0.6485
CPCosine	0.6274	0.6479	0.6524	0.6388	0.6479	0.6479	0.6520	0.6524	0.6467	0.6548	0.6568	0.6458	0.6457
CPLogistic	0.6241	0.6460	0.6302	0.6284	0.6460	0.6460	0.6306	0.6302	0.6497	0.6325	0.6538	0.6423	0.6272
CPCorpus	0.6282	0.6364	0.6256	0.6202	0.6364	0.6364	0.6258	0.6256	0.6505	0.6284	0.6591	0.6458	0.6390
IC models introduced in this work													
CPRefHyponyms	0.6327	0.6495	0.6510	0.6424	0.6495	0.6495	0.6510	0.6510	0.6530	0.6528	0.6598	0.6468	0.6452
CPRefUniform	0.6384	**0.6679**	0.6236	0.6366	**0.6679**	**0.6679**	0.6236	0.6236	0.6486	0.6262	0.6665	0.6525	0.6395
CPRefLeaves	0.6344	0.6526	0.6551	0.6429	0.6526	0.6526	0.6551	0.6551	0.6538	0.6553	0.6609	0.6470	0.6485
CPRefCosine	0.6318	0.6459	0.6515	0.6416	0.6459	0.6459	0.6515	0.6515	0.6531	0.6536	0.6598	0.6463	0.6368
CPRefLogistic	0.6323	0.6538	0.6490	0.6355	0.6538	0.6538	0.6490	0.6490	0.6639	0.6515	0.6639	0.6578	0.6539
CPRefCorpus	0.6341	0.6435	0.6284	0.6297	0.6435	0.6435	0.6284	0.6284	0.6534	0.6328	0.6617	0.6504	0.6390
CPRefCosineLeaves	0.6314	0.6485	0.6537	0.6437	0.6485	0.6485	0.6537	0.6537	0.6529	0.6567	0.6598	0.6463	0.6419
CPRefLogisticLeaves	0.6312	0.6554	0.6509	0.6380	0.6554	0.6554	0.6509	0.6509	0.6650	0.6524	0.6638	0.6589	0.6558
CPRefLeavesSubsumersRatio	0.6424	0.6558	0.6540	0.6474	0.6558	0.6558	0.6540	0.6540	0.6539	0.6541	0.6590	0.6494	0.6475
Best column value	0.6481	**0.6679**	0.6612	0.6656	**0.6679**	**0.6679**	0.6612	0.6612	0.6656	0.6619	0.6678	0.6662	0.6670

Table 20: Spearman (ρ) correlation coefficients for all the IC models and measures in the Agirre201 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

P&S$_{full}$ - r	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
IC models	Resnik	Lin	*J&C*	*P&S*	FaITH	Meng$_{12}$	Garla	*cosJ&C*	*Li*$_{s9}$	Zhou	Meng$_{14}$	Gao	*coswJ&C*
Resnik (1995)	0.8736	0.8877	0.8749	0.8328	0.9003	0.8988	0.7761	0.8922	0.8857	0.8778	0.8286	0.8909	0.8826
Seco et al. (2004)	0.8799	0.8945	0.8781	**0.8970**	0.9042	0.9031	0.9021	0.8966	0.8839	**0.8949**	0.8374	0.8585	0.8819
Blanchard et al. (2008)	0.8788	0.8932	0.8736	0.8381	0.9020	0.9013	0.7813	0.8942	0.8890	0.8769	**0.8375**	**0.8926**	0.8811
Zhou et al. (2008a)	0.8357	0.8420	0.8372	0.8563	0.8905	0.8806	0.8979	0.8726	**0.8897**	0.8725	0.7488	0.8585	0.8593
Sánchez et al. (2011)	0.8740	0.8738	0.8762	0.8051	0.9025	0.8964	0.8470	**0.8996**	0.8807	0.8789	0.7943	0.8844	**0.8850**
Sánchez and Batet (2012)	**0.8829**	**0.8948**	0.8742	0.8398	0.9042	0.9035	0.7798	0.8918	0.8877	0.8771	0.8350	0.8920	0.8790
Meng et al. (2012)	0.8645	0.8863	0.8715	0.8897	0.9057	0.9025	**0.9058**	0.8917	0.8813	0.8805	0.8106	0.8585	0.8765
Yuan et al. (2013)	0.8655	0.8896	0.8641	0.8891	**0.9082**	**0.9061**	0.9010	0.8840	0.8774	0.8796	0.8085	0.8585	0.8713
Hadj Taieb et al. (2014a)	0.4915	0.7962	0.4261	0.4660	0.8167	0.8125	0.6749	0.4272	0.8724	0.4308	0.7632	0.8751	0.4272
Adhikari et al. (2015)	0.8653	0.8875	0.8735	0.8908	0.9056	0.9027	0.9057	0.8926	0.8804	0.8819	0.8163	0.8585	0.8774
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.8725	0.8903	**0.8783**	0.8325	0.9015	0.9002	0.7884	0.8963	0.8865	0.8811	0.8286	0.8899	0.8808
CPUniform	0.7077	0.8042	0.7204	0.6602	0.8644	0.8498	0.7575	0.7478	0.8759	0.7260	0.7300	0.8696	0.8169
CPLeaves	0.8715	0.8887	0.8748	0.8297	0.9008	0.8992	0.7865	0.8930	0.8864	0.8775	0.8277	0.8897	0.8777
CPCosine	0.8757	0.8915	0.8778	0.8318	0.9015	0.9005	0.7502	0.8943	0.8795	0.8797	0.8330	0.8807	0.8785
CPLogistic	0.8419	0.8928	0.8315	0.8061	0.9064	0.9056	0.7742	0.8632	0.8852	0.8363	0.8193	0.8880	0.8645
CPCorpus	0.8649	0.8840	0.8711	0.8203	0.8987	0.8964	0.8129	0.8892	0.8858	0.8744	0.8246	0.8906	0.8807
IC models introduced in this work													
CPRefHyponyms	0.8690	0.8870	0.8730	0.8312	0.8978	0.8966	0.7852	0.8918	0.8892	0.8763	0.8353	0.8912	0.8808
CPRefUniform	0.7577	0.8738	0.7511	0.7301	0.8979	0.8942	0.7564	0.7771	0.8763	0.7560	0.7915	0.8758	0.8169
CPRefLeaves	0.8678	0.8854	0.8694	0.8282	0.8971	0.8955	0.7835	0.8887	0.8892	0.8726	0.8343	0.8908	0.8777
CPRefCosine	0.8694	0.8893	0.8748	0.8311	0.8992	0.8982	0.7953	0.8903	0.8883	0.8787	0.8339	0.8912	0.8748
CPRefLogistic	0.8403	0.8709	0.8564	0.8142	0.8923	0.8885	0.7977	0.8797	0.8827	0.8599	0.8347	0.8751	0.8608
CPRefCorpus	0.8631	0.8824	0.8687	0.8216	0.8957	0.8935	0.8072	0.8862	0.8888	0.8722	0.8325	0.8919	0.8807
CPRefCosineLeaves	0.8684	0.8864	0.8689	0.8263	0.8978	0.8962	0.7898	0.8856	0.8882	0.8726	0.8323	0.8909	0.8706
CPRefLogisticLeaves	0.8407	0.8695	0.8558	0.8130	0.8920	0.8878	0.7950	0.8788	0.8821	0.8593	0.8328	0.8743	0.8603
CPRefLeavesSubsumersRatio	0.8740	0.8874	0.8677	0.8286	0.8989	0.8974	0.7696	0.8808	0.8890	0.8699	0.8349	0.8910	0.8721
Best column value	0.8829	0.8948	0.8783	0.8970	**0.9082**	0.9061	0.9058	0.8996	0.8897	0.8949	0.8375	0.8926	0.8850

Table 21: Pearson (r) correlation coefficients for all the IC models and measures in the *P&S*$_{full}$ dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

P&S$_{full}$ - ρ	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
IC models	Resnik	Lin	*J&C*	*P&S*	FaITH	Meng$_{12}$	Garla	*cosJ&C*	*Li*$_{s9}$	Zhou	Meng$_{14}$	Gao	*coswJ&C*
Resnik (1995)	**0.7783**	0.7660	0.7717	0.7422	0.7660	0.7660	0.7717	0.7717	0.7809	0.7737	0.8105	**0.8089**	0.7773
Seco et al. (2004)	0.7735	0.7911	0.7768	0.7919	0.7911	0.7911	0.7768	0.7768	0.7889	0.8144	0.8150	0.7986	0.7821
Blanchard et al. (2008)	0.7527	0.7788	0.7668	0.7629	0.7788	0.7788	0.7668	0.7668	0.7793	0.7737	0.8121	0.7899	0.7751
Zhou et al. (2008a)	0.7720	0.7867	0.7999	0.7938	0.7867	0.7867	0.7999	0.7999	0.7903	**0.8212**	0.8122	0.7986	0.8005
Sánchez et al. (2011)	0.7732	**0.7936**	0.8003	0.7573	**0.7936**	**0.7936**	0.8003	0.8003	0.7789	0.8031	**0.8179**	0.8046	**0.8226**
Sánchez and Batet (2012)	0.7727	0.7854	0.7652	0.7681	0.7854	0.7854	0.7652	0.7652	0.7806	0.7732	0.8116	0.7958	0.7740
Meng et al. (2012)	0.7543	0.7776	**0.8127**	0.8122	0.7776	0.7776	**0.8127**	**0.8127**	0.7842	0.8126	0.8072	0.7986	0.8215
Yuan et al. (2013)	0.7759	0.7905	0.7957	**0.8199**	0.7905	0.7905	0.7957	0.7957	**0.7917**	0.8138	0.8125	0.7986	0.7975
Hadj Taieb et al. (2014a)	0.6140	0.7463	0.6094	0.6107	0.7463	0.7463	0.6094	0.6094	0.7674	0.6153	0.7833	0.7798	0.6094
Adhikari et al. (2015)	0.7541	0.7770	0.8068	0.8045	0.7770	0.7770	0.8068	0.8068	0.7842	0.8122	0.8066	0.7986	0.8164
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.7461	0.7824	0.7910	0.7790	0.7824	0.7824	0.7910	0.7910	0.7744	0.7943	0.8103	0.7885	0.7819
CPUniform	0.6997	0.7852	0.7684	0.6905	0.7852	0.7852	0.7684	0.7684	0.7804	0.7696	0.7973	0.7833	0.7974
CPLeaves	0.7457	0.7808	0.7766	0.7689	0.7808	0.7808	0.7766	0.7766	0.7745	0.7833	0.8089	0.7884	0.7818
CPCosine	0.7689	0.7834	0.7710	0.7664	0.7834	0.7834	0.7710	0.7710	0.7788	0.7760	0.8118	0.8053	0.7791
CPLogistic	0.7783	0.7921	0.7880	0.7630	0.7921	0.7921	0.7880	0.7880	0.7811	0.7888	0.8075	0.7910	0.7695
CPCorpus	0.7691	0.7572	0.7615	0.7291	0.7572	0.7572	0.7615	0.7615	0.7809	0.7656	0.8116	0.8077	0.7628
IC models introduced in this work													
CPRefHyponyms	0.7545	0.7692	0.7640	0.7603	0.7692	0.7692	0.7640	0.7640	0.7789	0.7682	0.8061	0.7909	0.7819
CPRefUniform	0.7189	0.7752	0.7394	0.7124	0.7752	0.7752	0.7394	0.7394	0.7763	0.7434	0.8026	0.7782	0.7974
CPRefLeaves	0.7528	0.7704	0.7654	0.7600	0.7704	0.7704	0.7654	0.7654	0.7784	0.7738	0.8047	0.7774	0.7818
CPRefCosine	0.7520	0.7836	0.7763	0.7641	0.7836	0.7836	0.7763	0.7763	0.7819	0.7800	0.8134	0.7913	0.7909
CPRefLogistic	0.7288	0.7335	0.7513	0.7113	0.7335	0.7335	0.7513	0.7513	0.7425	0.7616	0.7626	0.7342	0.7742
CPRefCorpus	0.7539	0.7522	0.7460	0.7247	0.7522	0.7522	0.7460	0.7460	0.7804	0.7506	0.8027	0.7925	0.7628
CPRefCosineLeaves	0.7520	0.7710	0.7672	0.7585	0.7710	0.7710	0.7672	0.7672	0.7832	0.7728	0.8078	0.7931	0.7841
CPRefLogisticLeaves	0.7291	0.7327	0.7595	0.7140	0.7327	0.7327	0.7595	0.7595	0.7445	0.7635	0.7609	0.7334	0.7822
CPRefLeavesSubsumersRatio	0.7578	0.7582	0.7677	0.7507	0.7582	0.7582	0.7677	0.7677	0.7787	0.7734	0.7999	0.7872	0.7804
Best column value	0.7783	0.7936	0.8127	0.8199	0.7936	0.7936	0.8127	0.8127	0.7917	0.8212	0.8179	0.8089	**0.8226**

Table 22: Spearman (ρ) correlation coefficients for all the IC models and measures in the *P&S*$_{full}$ dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

SimLex665 - r	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
IC models	Resnik	Lin	*J&C*	*P&S*	FaITH	Meng_{12}	Garla	*cosJ&C*	Li_{s9}	Zhou	Meng_{14}	Gao	*coswJ&C*
Resnik (1995)	0.5355	0.5935	0.5692	0.5489	0.5948	0.5955	0.4644	0.5781	0.6110	0.5753	0.5659	0.6152	0.5775
Seco et al. (2004)	0.5339	**0.6010**	0.5918	0.5975	0.6046	0.6048	0.6035	0.6013	0.5763	**0.6237**	0.5712	0.5602	0.6027
Blanchard et al. (2008)	0.5320	0.5958	0.5828	0.5517	0.5986	0.5990	0.4653	0.5935	**0.6133**	0.5882	0.5696	**0.6154**	0.5960
Zhou et al. (2008a)	0.5110	0.5778	0.5825	0.5803	0.5985	0.5945	0.6092	0.5973	0.5912	0.6162	0.5300	0.5602	0.6051
Sánchez et al. (2011)	**0.5364**	0.5791	0.5838	0.5318	0.5904	0.5883	0.5117	0.5941	0.6064	0.5887	0.5448	0.6112	0.5918
Sánchez and Batet (2012)	0.5354	0.5964	0.5861	0.5532	0.5991	0.5995	0.4637	0.5945	0.6124	0.5910	0.5683	0.6153	0.5976
Meng et al. (2012)	0.4972	0.5841	0.5939	0.5863	0.5887	0.5884	0.6064	0.6010	0.5661	0.6149	0.5526	0.5602	0.6056
Yuan et al. (2013)	0.5120	0.6004	**0.6027**	**0.5981**	**0.6063**	**0.6062**	**0.6166**	**0.6106**	0.5722	0.6203	0.5579	0.5602	**0.6157**
Hadj Taieb et al. (2014a)	0.2033	0.4849	0.2154	0.3237	0.4921	0.4896	0.3572	0.2177	0.5359	0.2192	0.5078	0.5448	0.2177
Adhikari et al. (2015)	0.4983	0.5848	0.5945	0.5871	0.5893	0.5890	0.6069	0.6016	0.5648	0.6155	0.5541	0.5602	0.6060
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.5223	0.5910	0.5811	0.5468	0.5932	0.5940	0.4573	0.5896	0.6118	0.5867	0.5660	0.6144	0.6023
CPUniform	0.3632	0.5245	0.5366	0.4693	0.5186	0.5260	0.3908	0.5416	0.6044	0.5408	0.5091	0.5992	0.6010
CPLeaves	0.5204	0.5904	0.5808	0.5457	0.5927	0.5934	0.4565	0.5897	0.6118	0.5864	0.5652	0.6143	0.6033
CPCosine	0.5274	0.5934	0.5841	0.5490	0.5959	0.5964	0.4169	0.5920	0.6063	0.5872	0.5688	0.6109	0.5984
CPLogistic	0.5068	0.5926	0.5679	0.5441	0.5966	0.5972	0.3876	0.5778	0.6098	0.5726	0.5589	0.6141	0.5976
CPCorpus	0.5271	0.5856	0.5564	0.5404	0.5844	0.5863	0.0593	0.5642	0.6112	0.5639	0.5630	0.6152	0.5759
IC models introduced in this work													
CPRefHyponyms	0.5204	0.5935	0.5864	0.5525	0.5962	0.5966	0.4689	0.5950	0.6133	0.5918	0.5704	0.6141	0.6023
CPRefUniform	0.4373	0.5704	0.5563	0.5180	0.5719	0.5728	0.4269	0.5612	0.6046	0.5599	0.5507	0.6050	0.6010
CPRefLeaves	0.5187	0.5931	0.5869	0.5519	0.5962	0.5965	0.4693	0.5958	0.6132	0.5921	0.5697	0.6138	0.6033
CPRefCosine	0.5207	0.5966	0.5875	0.5557	0.5978	0.5987	0.4718	0.5943	0.6128	0.5931	**0.5723**	0.6141	0.5941
CPRefLogistic	0.4940	0.5639	0.5744	0.5349	0.5781	0.5758	0.4519	0.5824	0.6039	0.5787	0.5522	0.6003	0.5895
CPRefCorpus	0.5268	0.5886	0.5611	0.5453	0.5883	0.5895	0.4682	0.5693	0.6132	0.5683	0.5677	0.6152	0.5759
CPRefCosineLeaves	0.5188	0.5963	0.5889	0.5548	0.5984	0.5990	0.4735	0.5964	0.6129	0.5943	0.5712	0.6139	0.5973
CPRefLogisticLeaves	0.4914	0.5636	0.5744	0.5343	0.5785	0.5760	0.4526	0.5828	0.6033	0.5786	0.5509	0.5996	0.5903
CPRefLeavesSubsumersRatio	0.5124	0.5908	0.5837	0.5455	0.5922	0.5931	0.4541	0.5912	0.6131	0.5878	0.5670	0.6139	0.6013
Best column value	0.5364	0.6010	0.6027	0.5981	0.6063	0.6062	0.6166	0.6106	0.6133	**0.6237**	0.5723	0.6154	0.6157

Table 23: Pearson (r) correlation coefficients for all the IC models and measures in the SimLex665 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

SimLex665 - ρ	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
IC models	Resnik	Lin	$J\&C$	$P\&S$	FaITH	Meng_{12}	Garla	$cosJ\&C$	Li_{s9}	Zhou	Meng_{14}	Gao	$coswJ\&C$
Resnik (1995)	0.5210	0.5810	0.5700	0.5450	0.5810	0.5810	0.5700	0.5700	0.5920	0.5741	0.6057	0.5994	0.5683
Seco et al. (2004)	0.5210	0.5888	0.5901	0.5862	0.5888	0.5888	0.5901	0.5901	0.5809	0.6101	**0.6105**	0.5918	0.5916
Blanchard et al. (2008)	0.5190	0.5844	0.5820	0.5466	0.5844	0.5844	0.5820	0.5820	0.5892	0.5853	0.6071	0.5955	0.5848
Zhou et al. (2008a)	0.5026	0.5841	0.5945	0.5767	0.5841	0.5841	0.5945	0.5945	0.5804	0.6089	0.6025	0.5918	0.6001
Sánchez et al. (2011)	**0.5248**	0.5803	0.5906	0.5204	0.5803	0.5803	0.5906	0.5906	0.5888	0.5933	0.6015	0.5953	0.5873
Sánchez and Batet (2012)	0.5232	0.5850	0.5835	0.5476	0.5850	0.5850	0.5835	0.5835	0.5896	0.5866	0.6068	0.5965	0.5868
Meng et al. (2012)	0.4896	0.5767	0.5957	0.5714	0.5767	0.5767	0.5957	0.5957	0.5672	0.6050	0.5964	0.5918	0.5980
Yuan et al. (2013)	0.5030	**0.5903**	**0.6027**	**0.5869**	**0.5903**	**0.5903**	**0.6027**	**0.6027**	0.5727	**0.6114**	0.6051	0.5918	**0.6076**
Hadj Taieb et al. (2014a)	0.3004	0.4833	0.3256	0.4243	0.4833	0.4833	0.3256	0.3256	0.5275	0.3307	0.5432	0.5423	0.3256
Adhikari et al. (2015)	0.4899	0.5766	0.5960	0.5718	0.5766	0.5766	0.5960	0.5960	0.5673	0.6050	0.5968	0.5918	0.5983
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.5067	0.5799	0.5806	0.5391	0.5798	0.5799	0.5723	0.5806	0.5873	0.5847	0.6046	0.5920	0.5922
CPUniform	0.3649	0.5223	0.5506	0.4620	0.5220	0.5223	0.5672	0.5506	0.5825	0.5530	0.5572	0.5846	0.6028
CPLeaves	0.5043	0.5797	0.5799	0.5392	0.5796	0.5797	0.5716	0.5799	0.5867	0.5839	0.6044	0.5924	0.5929
CPCosine	0.5150	0.5821	0.5828	0.5413	0.5821	0.5821	0.5745	0.5827	0.5898	0.5849	0.6070	0.5924	0.5884
CPLogistic	0.4925	0.5791	0.5738	0.5358	0.5790	0.5791	0.5738	0.5738	0.5885	0.5767	0.6032	0.5936	0.5904
CPCorpus	0.5085	0.5735	0.5578	0.5335	0.5735	0.5735	0.5460	0.5578	0.5912	0.5639	0.6024	0.5983	0.5662
IC models introduced in this work													
CPRefHyponyms	0.5077	0.5822	0.5856	0.5470	0.5822	0.5822	0.5856	0.5856	0.5896	0.5894	0.6076	0.5946	0.5922
CPRefUniform	0.4438	0.5599	0.5639	0.5097	0.5599	0.5599	0.5639	0.5639	0.5861	0.5662	0.5931	0.5914	0.6028
CPRefLeaves	0.5073	0.5817	0.5864	0.5469	0.5817	0.5817	0.5864	0.5864	0.5891	0.5901	0.6074	0.5943	0.5929
CPRefCosine	0.5081	0.5844	0.5864	0.5509	0.5844	0.5844	0.5864	0.5864	0.5879	0.5906	0.6099	0.5920	0.5863
CPRefLogistic	0.4876	0.5646	0.5763	0.5231	0.5646	0.5646	0.5763	0.5763	0.5878	0.5797	0.6007	0.5902	0.5803
CPRefCorpus	0.5084	0.5762	0.5612	0.5407	0.5762	0.5762	0.5612	0.5612	**0.5929**	0.5667	0.6045	**0.5998**	0.5662
CPRefCosineLeaves	0.5056	0.5840	0.5881	0.5498	0.5840	0.5840	0.5881	0.5881	0.5894	0.5915	0.6089	0.5927	0.5897
CPRefLogisticLeaves	0.4869	0.5657	0.5777	0.5235	0.5657	0.5657	0.5777	0.5777	0.5868	0.5808	0.6004	0.5897	0.5817
CPRefLeavesSubsumersRatio	0.5112	0.5785	0.5859	0.5466	0.5785	0.5785	0.5859	0.5859	0.5900	0.5884	0.6030	0.5952	0.5946
Best column value	0.5248	0.5903	0.6027	0.5869	0.5903	0.5903	0.6027	0.6027	0.5929	**0.6114**	0.6105	0.5998	0.6076

Table 24: Spearman (ρ) correlation coefficients for all the IC models and measures in the SimLex665 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

IC models	Resnik (1995)	Seco et al. (2004)	Blanchard et al. (2008)	Zhou et al. (2008a)	Sánchez et al. (2011)	Sánchez and Batet (2012)	Meng et al. (2012)	Yuan et al. (2013)	Hadj Taieb et al. (2014a)	Adhikari et al. (2015)	CPHyponyms	CPUniform	CPLeaves	CPCosine	CPLogistic	CPCorpus	CPRefHyponyms	CPRefUniform	CPRefLeaves	CPRefCosine	CPRefLogistic	CPRefCorpus	CPRefCosineLeaves	CPRefLogisticLeaves	CPRefLeavesSubsumersRatio
Resnik	—	.008	.496	.019	.122	.368	.061	.020	.000	.084	.420	.000	.038	.409	.000	.000	.422	.000	.414	.133	.000	.001	.332	.000	.011
Seco et al	.008	—	.000	.020	.149	.001	.036	.005	.000	.010	.000	.000	.000	.000	.000	.000	.002	.000	.001	.003	.000	.001	.001	.000	.001
Blanchard	.496	.000	—	.001	.172	.111	.024	.001	.000	.031	.387	.000	.001	.328	.000	.000	.375	.000	.289	.003	.000	.002	.211	.000	.006
Zhou et al	.019	.020	.001	—	.306	.003	.329	.173	.000	.053	.000	.000	.000	.000	.000	.000	.004	.000	.002	.007	.000	.001	.001	.000	.001
Sánchez$_{11}$	.122	.149	.172	.306	—	.219	.362	.256	.000	.444	.124	.000	.029	.135	.002	.001	.156	.000	.144	.297	.000	.006	.140	.000	.027
Sánchez$_{12}$	.368	.001	.111	.003	.219	—	.041	.002	.000	.058	.278	.000	.001	.170	.000	.000	.212	.000	.134	.041	.000	.001	.115	.000	.004
Meng et al	.061	.036	.024	.329	.362	.041	—	.159	.000	.010	.002	.000	.001	.010	.002	.001	.028	.000	.026	.079	.000	.003	.017	.000	.007
Yuan et al	.020	.005	.001	.173	.256	.002	.159	—	.000	.056	.001	.000	.000	.001	.000	.000	.005	.000	.004	.010	.000	.001	.002	.000	.001
H.Taieb	.000	.000	.000	.000	.000	.000	.000	.000	—	.000	.000	.001	.000	.000	.000	.000	.000	.000	.000	.000	.001	.000	.000	.001	.000
Adhikari	.084	.010	.031	.053	.444	.058	.010	.056	.000	—	.003	.000	.000	.013	.002	.001	.037	.000	.033	.118	.000	.003	.021	.000	.008
CPHypo	.420	.000	.387	.000	.124	.278	.002	.001	.000	.003	—	.000	.020	.465	.006	.003	.468	.001	.479	.096	.000	.019	.467	.000	.084
CPUnif	.000	.000	.000	.000	.000	.000	.000	.000	.001	.000	.000	—	.000	.000	.001	.001	.000	.021	.000	.000	.013	.001	.000	.010	.001
CPLeaves	.038	.000	.001	.000	.029	.001	.001	.000	.000	.000	.020	.000	—	.001	.013	.004	.013	.001	.005	.000	.000	.025	.010	.000	.218
CPCosine	.409	.000	.328	.000	.135	.170	.010	.001	.000	.013	.465	.000	.001	—	.001	.001	.479	.000	.499	.025	.000	.004	.407	.000	.027
CPLogistic	.000	.000	.000	.000	.002	.000	.002	.000	.000	.002	.006	.001	.013	.001	—	.092	.000	.002	.000	.000	.001	.418	.000	.006	.013
CPCorpus	.000	.000	.000	.000	.001	.000	.001	.000	.000	.001	.003	.001	.004	.001	.092	—	.000	.003	.000	.000	.002	.029	.000	.015	.003
CPRefHypo	.422	.002	.375	.004	.156	.212	.028	.005	.000	.037	.468	.000	.013	.479	.000	.000	—	.000	.459	.025	.000	.000	.295	.000	.001
CPRefUnif	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.021	.001	.000	.002	.003	.000	—	.000	.000	.110	.001	.000	.084	.000
CPRefLea	.414	.001	.289	.002	.144	.134	.026	.004	.000	.033	.479	.000	.005	.499	.000	.000	.459	.000	—	.007	.000	.001	.371	.000	.002
CPRefCos	.133	.003	.003	.007	.297	.041	.079	.010	.000	.118	.096	.000	.000	.025	.000	.000	.025	.000	.007	—	.000	.000	.001	.000	.001
CPRefLog	.000	.000	.000	.000	.000	.000	.000	.000	.001	.000	.000	.013	.000	.000	.001	.002	.000	.110	.000	.000	—	.002	.000	.009	.000
CPRefCorpus	.001	.001	.002	.001	.006	.001	.003	.001	.000	.003	.019	.001	.025	.004	.418	.029	.000	.001	.001	.000	.002	—	.001	.011	.006
CPRefCosLea	.332	.001	.211	.001	.140	.115	.017	.002	.000	.021	.467	.000	.010	.407	.000	.000	.295	.000	.371	.001	.000	.001	—	.000	.007
CPRefLogLea	.000	.000	.000	.000	.000	.000	.000	.000	.001	.000	.000	.010	.000	.000	.006	.015	.000	.084	.000	.000	.009	.011	.000	—	.001
CPRefLeSuRat	.011	.001	.006	.001	.027	.004	.007	.001	.000	.008	.084	.001	.218	.027	.013	.003	.001	.000	.002	.001	.000	.006	.007	.000	—

Table 25: Pairwise p-values of the one-sided t-Student distribution for the paired difference between the average Spearman (ρ) correlation values of each pair of IC models on all the IC-based similarity measures. Each pairwise p-value is computed using the vector of average Spearman correlation values of each IC model with each IC-based similarity measure (rows in table 11) as paired random sample set. For a level of significance of 5%, each p-value ≤ 0.05 denotes a statistically significant higher or lower performance between these two IC models.

Simi. measures	Lastra & García (2015), coswJ&C	Zhou et al (2008)	Lastra & García (2015), cosJ&C	P&S (2008)	Taieb et al (2014)	Gao et al (2015)	J&C (1997)	Meng & Gu (2012)	P&S FaITH (2010)	Lin (1998)	Li et al (2003) Strat. 3	Li_{s9} (2003)	Li et al (2003) Strat. 4	Sánchez et al (2012)	Meng et al (2014)	Al-Mubaid & Nguyen (2009)	Resnik (1995)	Pedersen et al (2007)	Leacock & Chodorow (1998)	Garla	Rada et al (1989)	Wu & Palmer (1994)
L&G, coswJ&C	—	.248	.026	.045	.067	.084	.026	.020	.020	.020	.029	.034	.042	.024	.178	.038	.003	.029	.029	.026	.029	.006
Zhou et al	.248	—	.034	.184	.023	.028	.034	.002	.002	.002	.002	.012	.015	.006	.149	.019	.020	.002	.002	.034	.002	.002
L&G, cosJ&C	.026	.034	—	.346	.207	.338	1.00	.023	.023	.023	.129	.083	.088	.042	.372	.188	.020	.129	.129	1.00	.129	.005
P&S	.045	.184	.346	—	.239	.340	.346	.114	.114	.114	.185	.124	.123	.079	.463	.226	.007	.185	.185	.346	.185	.010
H.Taieb	.067	.023	.207	.239	—	.294	.207	.499	.499	.499	.363	.016	.036	.001	.022	.413	.096	.363	.363	.207	.363	.000
Gao et al	.084	.028	.338	.340	.294	—	.338	.345	.345	.345	.077	.087	.015	.020	.029	.163	.095	.077	.077	.338	.077	.004
J&C	.026	.034	1.00	.346	.207	.338	—	.023	.023	.023	.129	.083	.088	.042	.372	.188	.020	.129	.129	1.00	.129	.005
Meng & Gu	.020	.002	.023	.114	.499	.345	.023	—	1.00	1.00	.346	.166	.132	.064	.104	.430	.055	.346	.346	.023	.346	.005
P&S, FaITH	.020	.002	.023	.114	.499	.345	.023	1.00	—	1.00	.346	.166	.132	.064	.104	.430	.055	.346	.346	.023	.346	.005
Lin	.020	.002	.023	.114	.499	.345	.023	1.00	1.00	—	.346	.166	.132	.064	.104	.430	.055	.346	.346	.023	.346	.005
Li_{s3} et al	.029	.002	.129	.185	.363	.077	.129	.346	.346	.346	—	.313	.132	.130	.034	.411	.124	1.00	1.00	.129	1.00	.013
Li_{s9} et al	.034	.012	.083	.124	.016	.087	.083	.166	.166	.166	.313	—	.168	.019	.006	.166	.145	.313	.313	.083	.313	.000
Li_{s4} et al	.042	.015	.088	.123	.036	.015	.088	.132	.132	.132	.132	.168	—	.443	.001	.069	.281	.132	.132	.088	.132	.007
Sánchez et al	.024	.006	.042	.079	.001	.020	.042	.064	.064	.064	.130	.019	.443	—	.001	.044	.226	.130	.130	.042	.130	.001
Meng et al	.178	.149	.372	.463	.022	.029	.372	.104	.104	.104	.034	.006	.001	.001	—	.018	.056	.034	.034	.372	.034	.001
Al-Mubaid	.038	.019	.188	.226	.413	.163	.188	.430	.430	.430	.411	.166	.069	.044	.018	—	.099	.411	.411	.188	.411	.004
Resnik	.003	.020	.020	.007	.096	.095	.020	.055	.055	.055	.124	.145	.281	.226	.056	.099	—	.124	.124	.020	.124	.203
Pedersen	.029	.002	.129	.185	.363	.077	.129	.346	.346	.346	1.00	.313	.132	.130	.034	.411	.124	—	1.00	.129	1.00	.013
L &Ch.	.029	.002	.129	.185	.363	.077	.129	.346	.346	.346	1.00	.313	.132	.130	.034	.411	.124	1.00	—	.129	1.00	.013
Garla & Brandt	.026	.034	1.00	.346	.207	.338	1.00	.023	.023	.023	.129	.083	.088	.042	.372	.188	.020	.129	.129	—	.129	.005
Rada et al	.029	.002	.129	.185	.363	.077	.129	.346	.346	.346	1.00	.313	.132	.130	.034	.411	.124	1.00	1.00	.129	—	.013
W & P	.006	.002	.005	.010	.000	.004	.005	.005	.005	.005	.013	.000	.007	.001	.001	.004	.203	.013	.013	.005	.013	—

Table 26: Pairwise p-values of the one-sided t-Student distribution for the paired difference between the Spearman (ρ) correlation values of each pair of similarity measures on all datasets. Each pairwise p-value is computed using the vector of average Spearman correlation values of each similarity measure on all datasets (rows in table 12) as paired random sample set. For a level of significance of 5%, each p-value ≤ 0.05 denotes a statistically significant higher or lower performance between these two similarity measures.