Engineering Applications of Artificial Intelligence 46 (2015) 140–153

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A novel family of IC-based similarity measures with a detailed experimental survey on WordNet

Juan J. Lastra-Díaz *, Ana García-Serrano

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

ARTICLE INFO

Article history:
Received 13 April 2015
Received in revised form
27 August 2015
Accepted 6 September 2015
Available online 1 October 2015

Keywords:
Ontology-based semantic similarity measures
IC-based measures
Semantic similarity
Intrinsic and corpus-based information content models
Jiang–Conrath distance
Semantic similarity on WordNet survey

ABSTRACT

This paper introduces a novel family of ontology-based similarity measures based on the Information Content (IC) theory, a detailed state of the art, a large experimental survey into ontology-based similarity measures on WordNet, and a new comparison between intrinsic and corpus-based IC models. Our experiments are based on our implementation of a large set of similarity measures, intrinsic and corpus-based IC models, which are evaluated on two known datasets and three different WordNet versions. The new measures are called *weighted Jiang–Conrath distance* (*wJ&Cdist*) *and similarity* (*wJ&Csim*), *cosine-normalized Jiang–Conrath similarity* (*cosJ&Csim*) and *cosine-normalized weighted Jiang–Conrath similarity* (*coswJ&Csim*). Two of our similarity measures outperform the state-of-the-art measures on the RG65 dataset, and one of them obtains the third overall score on all the datasets and evaluated WordNet versions. The cosine-normalized similarity measures are a non-linear normalization of the classic Jiang–Conrath (J&C) distance and the new *wJ&C* distance. On the other hand, the *wJ&C* distance is a generalization of the classic J&C distance which is based on the length of the shortest path between concepts within an IC-based weighted graph. Our measures are based on two not previously considered notions: (1) a generalization of the classic J&C distance to any type of taxonomy, based on an IC-based weighted graph derived from the conditional probabilities between child and parent concepts, and (2) a non-linear normalization function that converts the ontology-based semantic distances into similarity functions. Finally, the corpus-based IC models based on the Resnik method obtain rivaling results as regards the state-of-the-art intrinsic IC models, when they are used with some unexplored WordNet-based frequency files. Therefore, this latter fact allows us to reconsider some previous conclusions about the outperformance of the intrinsic IC models over the corpus-based ones.

1. Introduction and positioning

The ontology-based similarity measures have found many applications in natural language processing (NLP), information retrieval (IR), and bioengineering. For example, in IR the aim is to retrieve resources that are semantically related to a user query both defined as concept sets. In this context, the word-to-word similarity measures can be extended to compute the distance between bags of concepts, or weighted concepts and individuals, thus, they are a key component in estimating the closeness between a user query and the relevant info to be retrieved. This approach is followed in Lastra-Díaz (2014), where we introduce a novel ontology-based IR model called *Intrinsic Ontological Spaces*, which is based on a metric space defined by the *wJ&Cdistance* introduced herein and disclosed in Lastra-Díaz and García-Serrano (2014). In Chan et al. (2011), the authors introduce a concept-based IR model for biomedical documents based on an ontology-based vector model, in which the document weights are computed using a non-linear function of a truncated version of the length of the shortest path between concepts. Next, we describe other applications of the ontology-based similarity measures. In Sánchez et al. (2015), the authors introduce the notion of semantic variance (SV) as a means of evaluating the quality of any ontology, which is defined as the variance of the semantic distance function, as defined in Batet et al. (2011), between each concept and the root. In Yan et al. (2014), the authors introduce an ontology-based inventive problem solving method which is based on a short-text similarity measure derived from the measure in Lin (1998). In Patwardhan et al. (2003), the authors introduce a word sense disambiguation (WSD) method based on the distributional hypothesis and the use of ontology-based similarity measures to select the closest evocated concept between a disambiguated

* Corresponding author.
E-mail addresses: jlastra@invi.uned.es (J.J. Lastra-Díaz), agarcia@lsi.uned.es (A. García-Serrano).

http://dx.doi.org/10.1016/j.engappai.2015.09.006
0952-1976/

word and its neighboring words. In Mihalcea et al. (2006), the authors propose a text similarity measure based on the combination of an IDF weighting scheme with any ontology-based similarity measure, which is evaluated in a paraphrase detection (PD) task. In Cross and Hu (2011), the authors review the use of semantic similarity measures on the ontology alignment (OA) problem and introduce a semantic alignment quality measure based on the difference between the similarity measure between the concepts in the base ontology and their image in the target ontology. In Fiorini et al. (2015), the authors propose a semantic indexing method for biomedical documents based on similarity measures. In Couto and Pinto (2013) and Pesquita et al. (2009), the authors survey other applications of ontology-based similarity measures in bioengineering, such as the prediction of protein functions.

A semantic similarity measure is a binary function that given two input words computes their degree of similarity as perceived by a human being. Unlike the semantic relatedness between words, which includes other semantic co-occurrence relationships such as "part-of" or selectional preferences, the similarity measures are constrained to "is-a" relationships. The similarity measures can be roughly categorized into two families: ontology-based and corpus-based. An ontology-based semantic similarity measure is a binary function $sim : C \times C \rightarrow \mathbb{R}$ that approximates as much as possible the degree of similarity as perceived by a human being. In the latter expression, C is a concept set belonging to a taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$, which is defined by a partially ordered set $(C, \leq_C)$ and an overall supreme element $\Gamma \in C$ called the root. A word is represented by a set of concepts within a base ontology, and the similarity between words is defined as the highest similarity value of the Cartesian product between both concept sets. On the other hand, most of corpus-based similarity measures are based on the distributional hypothesis Harris (1981), which states that words in similar contexts tend to share similar meanings. Most of distributional measures define the word meanings as a function of their context and the type of co-occurrence relationships that needs to be captured. For example, the word contexts could be small n-gram windows, or larger contexts such as sentences, paragraphs or documents. Despite there being different methods to represent the word meanings (contexts), such as sets, vectors, probability distributions, and graph nodes, the most popular representations rely on vector space models (VSM) (Turney and Pantel, 2010, Section 2.2). For example, in Gabrilovich and Markovitch (2007) the authors introduce a semantic relatedness method to compute word and document similarity, called ESA, which represents the meaning of a word or text as a weighted vector of Wikipedia concepts (articles), whose weights are defined by the cosine score between the input text vector and each Wikipedia base vector.

1.1. Ontology-based similarity measures versus corpus-based

The main advantage of the ontology-based measures is that the logic relationships between concepts, especially the "is-a" relationships, are hand-coded within the ontologies. A second advantage of these measures is that they are defined by closed formulas that only require a taxonomy to be evaluated. Therefore, they can be easily implemented, although their computational cost depends on the size of the ontology and the complexity of the required algorithms. In contrast, a serious drawback of the ontology-based measures in open domain applications, like the Web, is their limited lexical coverage, and the cost of creating and updating wide coverage ontologies. On the other hand, the corpus-based measures mainly rely on the distributional hypothesis, and compute the degree of similarity using an indirect approach that relies on the statistical co-occurrence between word contexts. In addition to the "is-a" relationships, co-occurring words can encode other types of semantic relationships. Therefore, the corpus-based measures "can confuse similarity with relatedness" (Li et al., 2015, Section1). Moreover, "it is commonly considered that distributional measures can only be used to capture semantic relatedness" (Harispe et al., 2015, Section 2.5.2) and "they have traditionally performed poorly when compared to WordNet-based measures" (Mohammad and Hirst, 2012, p1) in the similarity assessment task. Another drawback of the corpus-based measures is that they are commonly based on a pipeline of NLP and IR algorithms. According to the complexity of the measure, it could require syntactic pattern extraction, POS tagging, WSD and further methods, as well as external services and resources, leading them to a high computational cost and replication complexity. In addition, the corpus-based measures exhibit the classic problems related to the corpus statistics, such as the difficulty in obtaining a well-balanced corpus for all words and their senses. On the other hand, the main advantage of the corpus-based measures is that they offer a broader lexical coverage. In summary, the ontology-based similarity measures are efficient, robust and easy to implement, whilst the corpus-based measures offer a broader lexical coverage.

The mainstream of the research in corpus-based similarity measures is the proposal for hybrid concept-based distributional measures, which integrate KBs or explicit "is-a" semantic networks to bridge the lack of well-defined semantic knowledge. For example, in Patwardhan and Pedersen (2006) the authors introduce a similarity and relatedness measure which relies on the gloss vector overlapping between the extended WordNet gloss vectors of two input concepts. In Mohammad and Hirst (2006) the authors propose 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. In Alvarez and Lim (2007), the authors introduce another hybrid distributional similarity measure that relies on the product of two taxonomical WordNet-based functions with a gloss overlapping factor, which are defined on a WordNet subgraph that includes "is-a" and "part-of" relationships. Finally, in Li et al. (2013) and Li et al. (2015), the authors introduce a hybrid distributional measure whose core idea is that the similarity computation relies on truly "is-a" relationships, unlike traditional corpus-based measures. The Li et al. (2015) measure is based on a general-purpose "is-a" semantic network derived from a large web-based corpus. The semantic network is defined by a set of triplets $(c, e, (P(e|c), P(c|e)))$, where c is a hypernym of e. The words are categorized as concepts or entities depending on their hypernym hyponym ratio. The context of the concepts is defined as a vector whose weights are the conditional probabilities $P(e|c)$ of their subsumed entities, whilst the entity vectors are defined in the opposite way. The similarity is defined as the cosine function between concept vectors, or entity vectors. The underlying idea of the Li et al. method is to use the overlap between the extension sets (subsumed entities) of the concepts as an estimation of their similarity.

1.2. Focusing on ontology-based similarity measures

The recent progress in concept-based distributional measures has proven that this approach offers a good tradeoff between precision and lexical coverage, especially for general-purpose domains such as the Web. However, these measures require the semantic annotation of a large corpus with a good coverage of all the concepts required, which is not always possible. In addition, some of these large corpuses could be not publicly available. On the other hand, we prove herein that the ontology-based similarity measures even exhibit some margin of improvement with a low computational cost that deserves to be studied. In addition, there