# A Domain Independent Semantic Measure for Keyword Sense Disambiguation

María G. Buey
everis / NTT Data
Zaragoza, Spain
maria.granados.buey@everis.com

Carlos Bobed
University of Zaragoza
Zaragoza, Spain
cbobed@unizar.es

Jorge Gracia
University of Zaragoza
Zaragoza, Spain
jogracia@unizar.es

Eduardo Mena
University of Zaragoza
Zaragoza, Spain
emena@unizar.es

## ABSTRACT

Understanding the user's intention is crucial in human-machine interaction. When dealing with text input, Word Sense Disambiguation (WSD) techniques play an important role. WSD techniques typically require well-formed sentences as context to operate, and predefined catalogues of word senses. However, such conditions do not always apply, such as when there is a need to disambiguate keywords from a query, or sets of tags describing any Web resource.

In this paper, we propose a keyword disambiguation method based on the semantic relatedness between words and ontological terms. Taking advantage of the semantic information captured by word embeddings, our approach maps a set of input keywords to their meanings within a given target ontology. We focus on situations where the available linguistic information is very scarce, hampering natural language based approaches. Experimental results show the feasibility of our approach without previous training for target domains.

## CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; *Lexical semantics*; Search methodologies; • **Information systems** → *Similarity measures*; Information retrieval query processing;

## KEYWORDS

Semantic Relatedness, Word Embeddings, Word Sense Disambiguation, Concept Linking

**ACM Reference Format:**
María G. Buey, Carlos Bobed, Jorge Gracia, and Eduardo Mena. 2021. A Domain Independent Semantic Measure for Keyword Sense Disambiguation. In *The 36th ACM/SIGAPP Symposium on Applied Computing (SAC '21), March 22–26, 2021, Virtual Event, Republic of Korea.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3412841.3442141

## 1 INTRODUCTION

In any information system which requires user interaction, being able to understand the user intention is a crucial requirement. In particular, being capable of disambiguating the input words is frequently the starting point of an interpretation process. Natural Language Processing (NLP) techniques that tackle disambiguation usually assume the presence of rich linguistic information. However, users are used to *keyword search queries* (a.k.a., *Web search queries*), sets of words which are a projection of the actual information need that they are expressing [4]. For example, a user could type "island Java" to look for any information related to the island of Java. *Keyword search queries* exhibit their own language structure [3]; thus, they require specific methods to disambiguate the words in such a scenario, where rich linguistic information might not be available.

Regarding word and meaning representation, recent advances in NLP have focused on different word embedding models [14], which are a set of language modeling and feature learning techniques where elements from a vocabulary are mapped into a vector space capturing their *distributional semantics* [9]. However, one of their main limitations is that the possible meanings of a word are combined into a single representation. Such a limitation has been tackled so far by: 1) representing individual meanings of words as distinct vectors in the space (e.g., sense2vec), and 2) more recently, by adopting *contextual word embeddings* [13]. Regarding our problem, the first approach allows us to control the represented meanings, but there exist scenarios where we cannot know all the different senses at training time. Contextual word embeddings models, the second approach, are proved to capture complex characteristics of word use such as polysemy. These models heavily rely on the structure of a sentence; however, this is not the case of keyword input scenarios where the user introduce a set of words without any evident structure or even in an arbitrary order. In particular, these models assign very different vectors to the same word when appearing in a well-formed sentence and in a keyword search even when they would have the same meaning. To illustrate this issue, Table 1 shows several examples for the words 'Java' and 'Apache' using BERT [8] contextual word embeddings.

In this paper, we propose a keyword disambiguation method which is based on the semantic relatedness between words and ontological terms. Our proposal maps a set of input words to their appropriate meanings in a given target ontology, extending our previous works on semantic relatedness [10] and disambiguation [11] to exploit the semantic information captured by word embeddings. Being completely decoupled from the target ontology makes our approach easily adaptable to any domain: it only requires a specific embedding model (unsupervised) trained for such domain, which is easy to obtain from a domain document corpus. To evaluate the performance and flexibility of our approach, we have carried out a thorough experimentation in the context of Word Sense Disambiguation (WSD) in general domains, and Concept Linking in a more specific domain (clinical knowledge).

The rest of the paper is structured as follows. Section 2 discusses related works. In Section 3 we describe our semantic relatedness measure approximation. In Section 4 we present the disambiguation algorithm that we use, and Section 5 focuses on our experimental results. Finally, our conclusions and future work appear in Section 6.


*SAC '21, March 22–26, 2021, Virtual Event, Republic of Korea*

ACM ISBN 978-1-4503-8104-8/21/03.
https://doi.org/10.1145/3412841.3442141

**Table 1: Cosine distance between the vector of a word obtained from a natural language sentence and the vector for the same word in an equivalent keyword-based sentence (BERT contextualized embeddings were used).**

| Natural Language sentence | Keywords | Focus on word | Word distance |
|---|---|---|---|
| I want to visit the island of Java | island Java | Java | 0.23 |
| I am learning the Java Programming Language | Java Programming Language | Java | 0.09 |
| Yesterday, an Apache Attack produced a long server shutdown | Apache Attack | Apache | 0.34 |
| The Apache tribe attacked the fort that night | Apache Attack | Apache | 0.32 |

## 2 RELATED WORK

***Semantic Relatedness and WSD.*** Semantic relatedness is the degree in which two objects are related by any kind of semantic relationship [5] and lies at the core of many applications in NLP (such as WSD, or Concept Linking).

Regarding WSD methods, supervised and knowledge-based approaches are typically used. Supervised approaches make use of sense-annotated training data and exploit linguistic features from corpora as training information. However, one important drawback is that they strongly depend on a sense annotated corpora, which might not be available, and they need to be updated as the ontology evolves. Moreover, a target word or any of its senses may never be observed during training, and the system will not be able to annotate it. On the other hand, knowledge-based systems exploit linguistic properties of lexical resources to perform WSD. They usually create a graph representation of the input text to then exploit different graph analysis algorithms over it (e.g., PageRank). To the best of our knowledge the two SOTA knowledge-based systems are UKB [1] and KEF [16]. However, they heavily depend on generic lexical relationships that are difficult to find in general ontologies.

***Semantics and Embeddings.*** Depending on how they model meaning and where they obtain it from, embedding techniques providing meaning-aware word vectors can be classified in: 1) *contextual word embeddings* [13], which are unsupervised models that induce word senses from huge text corpora by analyzing their contextual semantics[1], and 2) knowledge-based methods which exploit sense inventories of lexical resources. Unfortunately, with contextual word embeddings, we cannot target a particular ontology as we do not have control over the concept detail/granularity, and the learned senses might not be aligned to any particular human-readable structure. In addition, they depend heavily on sentence structure, which render them not suitable for keyword inputs, where word omission and order alterations are frequent. Regarding knowledge-based embedding methods, they require to know all the senses at training time, not being easily adaptable to new scenarios (e.g., addition/deletion of senses, evolving ontologies, etc.). Thus, we aim at requiring neither re-training nor newly labelled data, while being capable to disambiguate words against any sense repository.

## 3 SEMANTIC RELATEDNESS MEASURE

Word embeddings can be used directly to compute relatedness between words. However, they do not suffice when ontological terms come into play. To calculate the semantic relatedness between a keyword and an ontological term, we ground on the relatedness measure proposed in [10] which focuses on the relatedness between words that appear in keyword-based inputs. Our cornerstone is the notion of *ontological context* of an ontological term (denoted by "$OC(t)$"), defined as the minimum set of other ontological terms that belong to its semantic description, locating the term in the ontology and characterizing its meaning. Thus, given an ontological term $t$ and a word $w$, we compute their relatedness measure as:

$$sRel(t,w) = \lambda\ sRel_0(t,w) + (1-\lambda)\ sRel_1(t,w),\ (0 \le \lambda \le 1) \quad (1)$$

$$sRel_0(t,w) = agg_0(\{rel_w(syn_{ti},w) | syn_{ti} \in Syn(t)\}) \quad (2)$$

$$sRel_1(t,w) = agg_1(\{sRel_0(oc_{ti},w) | oc_{ti} \in OC(t)\}) \quad (3)$$

with $sRel_0(t,w)$ measuring the relatedness of different synonyms of $t$ and $w$, $sRel_1(a,b)$ measuring the relatedness of $OC(t)$ and $w$, and $\lambda$ being a parameter which governs how their values are blended. $rel_w$ is the relatedness between words; $Syn(t) = \{syn_{t1}, syn_{t2}, ...\}$ are the synonyms (equivalent labels) of $t$; $OC(t) = \{oc_{t1}, oc_{t2}, ...\}$ are the terms of the ontological context of $t$[2]; and $agg_0$ and $agg_1$ are the aggregation functions applied to the sets of $rel_w$ and $sRel_0$ values, respectively (we advocate to use *average* or *maximum* functions, see Section 5)[3]. Thus, the ontological term is characterized by considering two levels of its semantic description: the term label and its synonyms (Equation 2), and its ontological context (Equation 3).

The original proposal for $rel_w$ in [10] was the Normalized Web Distance *NWD(x,y)*, a generalization of the Cilibrasi and Vitányi's Normalized Google Distance *NGD(x,y)* [7] to use any Web search engine as source of frequencies. NWD ranges from 0 to $\infty$, to obtain a relatedness measure in the range $[0,1]$ increasing inversely, the following transformation was applied:

$$rel_w(x,y) = relWeb(x,y) = e^{-2NWD(x,y)} \quad (4)$$

To exploit word embeddings, substituting $relWeb(x,y)$ by the *cosine distance*, as broadly adopted, was not possible as its values range in $[-1,1]$. So, to obtain measure in the range $[0,1]$, we propose to use the *angular distance* instead, which is computed as follows:

$$rel_w(x,y) = ang.distance(x,y) = 1 - \frac{arccos(sim(x,y))}{\pi} \quad (5)$$

Thus, we substitute Equation 4 by Equation 5 directly in Equation 2 (we validate this substitution experimentally in Section 5.1). For those cases in which the label of the ontological term is multi-word, we compute the centroid as it is broadly adopted.

## 4 KEYWORD DISAMBIGUATION

Extending the distributional hypothesis [9], our hypothesis is that the most significant words in the disambiguation context are the most highly related to the word to disambiguate; such words conform its *active context*, $C_a$. More formally, let $k$ be an element of an input sequence of words $\mathbb{S}$, $\mathbb{K} \subseteq \mathbb{S}$ be the set of all possible keywords

[1] We refer the interested reader to [13] for a complete survey on these models.

[2] Notice that $|Syn(x)| \ge 1$ and $|OC(x)| \ge 0$.

[3] In the original formulation, the average was proposed, but we generalize it to explore the influence of the linkage used between the sets.

in the input, $C \subseteq \mathbb{K}$ the set of keywords of the disambiguation context, and $k_d \in \mathbb{K}$ the target keyword to disambiguate. Thus:

Definition 1. *Given a context $C \in \mathbb{K}$, and a keyword to disambiguate $k_d \in \mathbb{K}$, the active context $C_a$ of $k_d$ is a subset $C_a \subseteq C$ such that* $\forall k_i \in C_a, \nexists k_j \in (C - C_a) \ni rel_w(k_j, k_d) > rel_w(k_i, k_d)$.

To obtain the *active context* $C_a \subseteq C$ of $k_d$, we: 1) remove repeated words and stopwords from $C$, 2) apply the semantic relatedness ($rel_w$ in our case) between each keyword $k_i \in C$ and $k_d$, and 3) construct $C_a$ with $k_i \in C$ whose relatedness scores above a certain threshold[4].

**Disambiguating the keywords**. We ground on our algorithm proposed in [11]. This algorithm[5] takes as input $k_d$, $C_a$, and a set of possible senses for $k_d$, $S_{k_d}$, and performs three main steps: 1) obtaining the semantic relatedness between $C_a$ and each candidate sense $s_i \in S_{k_d}$, 2) calculating the overlap between $C_a$ and $OC(s_i)$ for each $s_i \in S_{k_d}$, and 3) re-ranking $S_{k_d}$ according to their frequency of use (when such information is available). The output is a score for each sense $s_i \in S_{k_d}$ that represents the confidence level of being the right sense. Note that $S_{k_d}$ is not restricted to any particular dictionary, as it could be dynamically built from, e.g., different ontological resources.

**Algorithm extensions**. Apart from using the updated $rel_w$ within the relatedness formulae, we modify the second step to exploit word embeddings-based representations instead of using the overlap between the *active context* of the keyword being disambiguated and the *ontological context of a term* . We have studied the following methods to capture the influence of the contexts:

***Average:*** This method calculates the average vector of the different bag of words involved in the disambiguation, under the assumption that the semantically coherent groups of words should outstand from the others. Thus, this method computes the average relatedness between the word vectors from $C_a$ and $OC(s_i)$.

***Smooth Inverse Frequency (SIF):*** Arora et al. [2] proposed to represent a sentence by a weighted average vector of its word vectors, from which the most frequent component obtained using PCA/SVD is substracted. This component may encompass words that occur frequently in a corpus and lack semantic content, thus not contributing to the disambiguation. We propose to compute a new score for each sense in $S_{k_d}$ by measuring the distance between the centroid of the active context $C_a$ and the SIF vector of the $OC(s_i)$.

***Top-K Nearest Words:*** We apply the same *active context* hypothesis to *OCs*: the words in the *OC* of a sense which are the closest ones to $C_a$ and $k_d$ should be the most significant for the disambiguation. Thus, from each $OC(s_i)$, we select the top-$k$ nearest keywords to $C_a \cup k_d$. Then, we compute the distance between the centroids of $C_a$ and of the top-$k$ keywords of $OC(s_i)$ to obtain a new score.

We explored the performance of these methods to rule out non appropriate ones. We report the results obtained in the next section.

# 5 EXPERIMENTAL EVALUATION

In order to validate our proposal, we have carried out three main sets of experiments. Due to the lack of space, we only present here the overall conclusions obtained from the results[6].

## 5.1 Correlation with Human Judgment

We analysed the correlation of the *angular distance* with human judgment in a basic word-to-word comparison. For this purpose, we used different datasets available in the literature (see the summary presented by Lastra et. al [12]) and we used 12 different pre-trained word embeddings built with different techniques.

***Results:*** In general, using the *angular distance* to calculate relatedness between pairs of words offers semantic correlation with the human judgment. It varies depending on the dataset and the model used, but it achieves an average 59.79% of Spearman correlation (SD 17.73), whereas cosine distance achieves an average of 61.82% (SD 16.58). Thus, despite a small decrease of correlation with human judgment compared to the cosine distance (broadly adopted in the literature), these results allow us to use this measure for our purposes.

## 5.2 Word Sense Disambiguation Evaluation

To evaluate our proposal in a general domain, we used two datasets: SemEval 2013, and SemEval 2015, and WordNet as target ontology. As these datasets contain sentences in natural language and our proposal is focused on keyword-based inputs, we built two additional setups: 1) starting from each natural language dataset, we derived a dataset by keeping the words of the most usual types in keyword expressions (noun, adjectives, and verbs)[7], and 2) to restrict even further the input, we derived a dataset from each of these latter ones by taking groups of three consecutive keywords. Regarding embeddings, we used the $NASARI_{embed}+UBMCw2v$ [6] vectors since in preliminary tests we saw that they offered the best results. Finally, we built the *OC* for each concept including its synonyms, hypernyms, and hyponyms. Besides, we also included their glosses (available in WordNet). The best configuration was achieved using the *average* aggregation function, and the *Top-K Nearest Words* method. We also witnessed that giving more importance to the *OC* improved the results ($\lambda = 0.25$).

***Results:*** We compare our best configuration with the current SOTA systems in WSD, specifically with: the supervised system proposed by Vial et al. [15], UKB [1], and KEF [16]. Table 2 presents the results obtained for all the systems.

Vial et al. [15] is compromised as we transform the input into a keyword expression. Although we did not outperformed its results, we are close in the SemEval 2013 dataset. Note that this system requires an annotated dataset, which might not be available in the target domain. Regarding UKB, it suffers when dealing with short inputs, where it does not have enough information. In such a setup, our proposal gets better results in SemEval 2013, and close ones in SemEval 2015. Besides, note that this approach strongly depends on the availability of lexical relationships, difficult to find in non-lexical domain ontologies. Regarding KEF, it remains stable in any case. However, this comes at the cost of additional computational cost[8] and, as UKB does, it heavily depends on lexical knowledge, so it is not portable to other domains with different ontologies. To sum up, our proposal achieves good performance in general domain scenarios where the linguistic information is scarce. In addition, it provides flexibility to work independently of the resources used.

## 5.3 Concept Linking Evaluation

To test the domain flexibility of our proposal, we performed an evaluation in the task of *Concept Linking* in the Health domain. For this, we used the dataset provided in the *Task 1* of the *ShaRE/CLEF eHealth*

[4] The maximum cardinality of $C_a$ is set to 4 following the suggestions in [11].
[5] We refer the interested reader to [11] for the complete details.
[6] The list of models (with their references), the datasets features, and the complete experimental results can be found at http://sid.cps.unizar.es/projects/kwdDisambiguation/
[7] Following the analysis done in [3], where they showed that 80% of the words used in Web searches belonged to that categories.
[8] In average, KEF took 1.2 s. per annotation, while our approach took 0.15 s. This difference is noticeable when we consider a complete sentence: for example, in SemEval 2015, with an average of 7.4 annotations per sentence, we would have 8.82 s. for KEF compared to 1.08 s. for our approach.

**Table 2: F-Score results moving from natural language sentences to only considering most used words in keyword expressions (noun, adjectives, and verbs), and to considering groups of three keywords.**

| System\Dataset | SemEval 2013 | | | SemEval 2015 | | |
|---|---|---|---|---|---|---|
| | Natural Language | Nouns & Adjs & Verbs | 3 Keywords | Natural Language | Nouns & Adjs & Verbs | 3 Keywords |
| *Vial et al. [15]* | 78.7 | 74.3 | 61.9 | 82.6 | 77.8 | 65.8 |
| *UKB* | 67.1 | 67.1 | 56.6 | 69.9 | 69.9 | 63.4 |
| *KEF* | 68.4 | 68.4 | 67.6 | 72.3 | 72.3 | 71.3 |
| *Our Proposal* | 64.7 | 65.4 | 60.6 | 60.1 | 59.3 | 58.1 |

**Table 3: Precision results in *Task 1* of the *ShaRE/CLEF eHealth 2013 Evaluation Lab.***

| Precision\Method | Top-K Nearest Words | | | ElasticSearch |
|---|---|---|---|---|
| | $\lambda = 0.25$ | $\lambda = 0.5$ | $\lambda = 0.75$ | |
| Precision | 70.62% | 76.05% | 78.78% | 66.73% |
| Precision@3 | 89.45% | 90.57% | 91.30% | 87.18% |

*2013 Evaluation Lab*. We addressed the *subtask b* which consists of mapping annotated disorder mentions to SNOMED-CT concepts included in UMLS. We used ElasticSearch[9] to index all the concepts which gave us a syntactic baseline to compare to. In this setup, we used all the mentions appearing in each document as the context of the mention to be disambiguated. For each mention, we retrieved N (set to 10) candidate concepts from ElasticSearch and we ran our disambiguation method. As word embeddings, we used a publicly available w2v model[10] trained on PMC&PubMed corpus.

***Results:*** The best method was *Top-K Nearest Words*, along with the *maximum* aggregation function. In this case, hypernyms did not contribute as much, and the best results were obtained when the OC contained synonyms and hyponyms. Table 3 reports the *precision* (P) and the *precision at 3* (P@3) results achieved. Ranking semantically the candidate concepts improved strongly the syntactic baseline results. We also noted the increasing performance in this scenario as we increased $\lambda$: in this ontology, concepts are very close to each other both syntactically and semantically, so, it is more important to give more weight to synonymy. Summing up, our proposal improves this concept linking task by helping disambiguating the terms in a different domain without any particular training for that, which shows the flexibility of our approach.

## 6 CONCLUSIONS AND FUTURE WORK

In this paper we have presented a keyword disambiguation approach based on a semantic relatedness measure which exploits the semantic information captured by word embeddings, capable of mapping words to their meanings within a given ontology. With our proposal:

- We provide a method to calculate the semantic relatedness between words and concepts of an ontology.
- We are able to disambiguate keyword-based inputs, where the linguistic information is scarce, and link them to concepts from an ontology in a flexible way. Our proposal can be adapted to any domain in a dictionary-decoupled way, lowering the potential data requirements (no annotated data is required).
- We have evaluated our proposal via thorough experimentation in general and specific domains, competing with current SOTA approaches and showing the flexibility of our approach.

As future work, we want to explore how contextual word embeddings [13] could be used in this context. Moreover, given the existing differences between general and specific domains (where concepts are both syntactic and semantically closer), we want to explore how we could take into account the syntactic and semantic features of the concepts to adapt dynamically our proposal to the scenario tackled.

[9]https://www.elastic.co/es/elasticsearch
[10]http://bio.nlplab.org/#word-vectors

## ACKNOWLEDGMENTS

This work has been supported by the European Union's Horizon 2020 projects Lynx (grant agreement No 780602) and Prêt-à-LLOD (grant agrement No 825182), and by Spanish national projects TIN2016-78011-C4-3-R (AEI/ FEDER, UE) and DGA/FEDER.

## REFERENCES

[1] E. Agirre, O. López de Lacalle, and A. Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics* 40, 1 (2014), 57–84.
[2] S. Arora, Y. Liang, and T. Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *Proc. of Intl. Conf. on Learning Representations (ICLR'17)*. 1–16.
[3] C. Barr, R. Jones, and M. Regelson. 2008. The Linguistic Structure of English Web-Search Queries. In *Proc. of Conf. on Empirical Methods in Natural Language Processing (EMNLP'08)*. 1021–1030.
[4] C. Bobed and E. Mena. 2016. QueryGen: Semantic interpretation of keyword queries over heterogeneous information systems. *Information Sciences* 329 (2016), 412–433.
[5] A. Budanitsky and G. Hirst. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics* 32, 1 (2006), 13–47.
[6] J. Camacho-Collados, M. T. Pilehvar, and R. Navigli. 2016. NASARI: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence* 240 (2016), 36–64.
[7] R. L. Cilibrasi and P. M. B. Vitanyi. 2007. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering* 19, 3 (2007), 370–383.
[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19)*. 4171–4186.
[9] J. R. Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* (1957).
[10] J. Gracia and E. Mena. 2008. Web-based Measure of Semantic Relatedness. In *Proc. of Intl. Conf. on Web Information Systems Engineering (WISE'08)*. 136–150.
[11] J. Gracia and E. Mena. 2009. Multiontology Semantic Disambiguation in Unstructured Web Contexts. In *Proc. of Workshop on Collective Knowledge Capturing and Representation (CKCaR'09) at K-CAP'09*. 1–9.
[12] J. J. Lastra-Díaz, J. Goikoetxea, M. A. H. Taieb, A. García-Serrano, M. B. Aouicha, and E. Agirre. 2019. A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art. *Engineering Applications of Artificial Intelligence* 85 (2019), 645–665.
[13] Qi Liu, Matt J. Kusner, and Phil Blunsom. 2020. A Survey on Contextual Embeddings. *arXiv preprint arXiv:2003.07278* (2020).
[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. 3111–3119.
[15] L. Vial, B. Lecouteux, and D. Schwab. 2019. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proc. of the Global WordNet Conference (GWC'19)*.
[16] Yinglin Wang, Ming Wang, and Hamido Fujita. 2020. Word sense disambiguation: A comprehensive knowledge exploitation framework. *Knowledge-Based Systems* 190 (2020), 105030.