EPiC Series in Computing

Volume 71, 2020, Pages 19–29

Vampire 2018 and Vampire 2019.
The 5th and 6th Vampire Workshops

Towards Word Sense Disambiguation by Reasoning

Javier Álvez[1], Itziar Gonzalez-Dios[2], and German Rigau[3]

[1] LoRea Group, University of the Basque Country (UPV/EHU)
javier.alvez@ehu.eus
[2] Ixa Group, University of the Basque Country (UPV/EHU)
itziar.gonzalezd@ehu.eus
[3] Ixa Group, University of the Basque Country (UPV/EHU)
german.rigau@ehu.eus

Abstract

In this paper, we describe a practical application of Vampire for Word Sense Disambiguation (WSD), which is an important research area in the field of Natural Language Processing (NLP). Its objective is choosing the intended sense of a word in a given context. In particular, we propose a method for the automatic disambiguation of the semantic relations in BLESS, which is a dataset designed to evaluate models of distributional semantics, by choosing the WordNet synset it belongs to. For this purpose, we use the knowledge in Adimen-SUMO, which is obtained by means of a suitable transformation of the knowledge in the core of SUMO[1] into first-order logic (FOL) and enables its use by FOL automated theorem provers such as Vampire. By exploiting the semantic mapping between WordNet and SUMO, we apply a black-box testing method that enables the automatic creation a set of conjectures for each word pair by considering the semantic relations provided by BLESS. Then, these conjectures are evaluated using Vampire and, according to the outcomes, each word is disambiguated to a single synset. Finally, we compare the results provided by our proposal and different disambiguation systems that can be found in the literature.

1 Introduction

Word Sense Disambiguation (WSD) [2] is an important research area in the field of Natural Language Processing (NLP). Several editions of the SemEval (Semantic Evaluation) series[2] include WSD tasks. The purpose of WSD is choosing the intended sense of a word in a given context. To that end, words are mapped to their corresponding WordNet synsets [14]. WordNet is a large lexical database of English where nouns, verbs, adjectives and adverbs are grouped into sets of synonyms (synsets). Each synset denotes a distinct concept and they are interlinked among them by means of lexical-semantic relations such as *synonymy*, *antonymy*, *hyponymy*, *meronymy* or morphosemantic relations.

In this paper, we describe a practical application of Vampire [19] for WSD. In particular, we propose a method for the automatic disambiguation of the semantic relations in BLESS [11],

[1] http://www.ontologyportal.org
[2] https://aclweb.org/aclwiki/SemEval_Portal

Sense	Context
"*A sweater or jersey with a high close-fitting collar*"	"*The black fleece is soft as cotton candy and is second on my preferred* ***turtle*** *list*"
"*Any of various aquatic and land reptiles having a bony shell and flipper-like limbs for swimming*"	"*Some* ***turtles*** *lay eggs in the sand and leave them to hatch on their own. The young* ***turtles*** *make their way to the top of the sand and scramble to the water while trying to avoid predators*"

Table 1: Senses of the word *turtle* in different contexts

which is a dataset designed to evaluate models of distributional semantics. It includes a set of tuples on different semantic relations, enabling the assessment of the ability of a model to detect related word pairs.

The task of WSD is the following: words can different meanings in different contexts. For example, the word *turtle* may refer to a sweater or a reptile, as described in Table 1. Therefore, given a word, we need to choose which is its sense, that is, which synset it belongs to. To that end, we use the knowledge in Adimen-SUMO [8], which is obtained by means of a suitable transformation of the knowledge in the core of SUMO[3] [20] into first-order logic (FOL) and enables its use by FOL automated theorem provers such as Vampire.

WordNet and SUMO (and therefore Adimen-SUMO) are connected in a semantic mapping [21] by means of three semantic relations: *equivalence*, *subsumption* and *instance*. By exploiting this mapping, we apply the black-box testing method described in [10] that enables the automatic creation a set of conjectures for each word pair by considering the semantic relations provided by BLESS. Then, these conjectures are evaluated using Vampire and, according to the results, each word is disambiguated to a single synset.

We have compared our proposal with UKB [1], which is one of the best WSD that can be found in the literature. Further, we have also evaluated the performance of these WSD by using a gold standard manually developed by a human expert. From this evaluation, we can conclude that Adimen-SUMO and Vampire can be successfully applied to NLP tasks.

Outline. In the next section, we introduce the main knowledge resources that are used in this paper. Then, we describe our WSD proposal in Section 3. Next, we report on the obtained experimentation results on 4. Finally, we provide some conclusions and discuss future work in Section 5.

2 Knowledge Resources

In this section, we introduce the knowledge resources that are used in our proposal, which are: i) BLESS; ii) WordNet; iii) SUMO and its FOL transformation Adimen-SUMO; and iv) the mapping between WordNet and SUMO.

BLESS data (*Baroni-Lenci Evaluation of Semantic Similarity*) [11] is a dataset designed for the evaluation of distributional semantic models. It includes 200 concrete nouns —called *targets*— (100 animate and 100 inanimate nouns) from different classes (e.g., *tools*, *clothing*, *vehicles*, *animals*, etc.). Each target is associated to a set of other words (nouns, verbs or adjectives) via six relations: *hypernymy*, *cohyponymy*, *meronymy*, *attribute*, *event* and *random*. In Table 2, we provide some examples of BLESS pairs involving the word *turtle*.

[3] http://www.ontologyportal.org

Relation	Pair
hypernymy	*turtles* are *amphibians*
cohyponymy	*turtles* and *frogs* are coordinate
meronymy	*turtles* have *legs*
attribute	*turtles* are *slow*
event	*turtles walk*

Table 2: Some BLESS pairs involving *turtle*

WordNet [14] is a large lexical database where nouns, verbs, adjectives and adverbs are grouped into sets of synonyms (*synsets*), each expressing a distinct concept. Each synset refers to a word sense using the following format: $word^s_p$, where s is the sense number and p is the part-of-speech (n for nouns, v for verbs and a for adjectives). Although superficially resembling a thesaurus, WordNet interlinks not just word forms but specific senses of words. Thus, the main relation in WordNet is synonymy, but synsets are interlinked by means of many conceptual-semantic and lexical relations such as the super- and subordinate relations hyperonymy and hyponymy.

SUMO[4] [20] is an upper level ontology proposed as a starter document by the IEEE Standard Upper Ontology Working Group. SUMO is expressed in SUO-KIF (Standard Upper Ontology Knowledge Interchange Format [22]), which is a dialect of KIF (Knowledge Interchange Format [15]). The syntax of both KIF and SUO-KIF goes beyond FOL and, therefore, SUMO axioms cannot be directly used by FOL ATPs without a suitable transformation. In [8], we proposed a translation of around an 88 % of the knowledge in the two upper levels of SUMO into FOL. As result, we obtained Adimen-SUMO, which was developed under the *Open World Assumption* (OWA) [13] and is currently included in the *Thousands of Problems for Theorem Provers* (TPTP) problem library[5] [26]. In this paper, we use an evolved version of Adimen-SUMO developed under the *Closed World Assumption* (CWA) [3]. The resulting ontology can be used in tasks that involve reasoning with commonsense knowledge. For example, our ontology includes the following axiom

```
(forall (?LIMB)
    (=>
        (instance ?LIMB Limb)
        (exists (?VERT)
            (and
                (instance ?VERT Vertebrate))
                (properPart ?LIMB ?VERT))))
```

which states that every instance of $Limb_c$ is part of some instance of $Vertebrate_c$.

WordNet is linked with SUMO by means of the mapping described in [21]. This mapping connects synsets of WordNet to terms of SUMO using three relations: *equivalence*, *subsumption* and *instantiation*. The relation *equivalence* denotes that the related WordNet synset and SUMO concept are equivalent in meaning, whereas *subsumption* and *instantiation* indicate that the WordNet synset is subsumed by the SUMO concept or is an instance of the SUMO concept respectively. We denote mapping relations by concatenating the symbols '=' (*equivalence*), '+' (*subsumption*) and '@' (*instantiation*). For example, the synsets $turtle^1_n$ and $turtle^2_n$ are respectively connected to $Clothing_c+$ and $Reptile_c+$ via *subsumption*, while $clothing^1_n$ and $Snake^1_n$ ("*a*

[4] http://www.ontologyportal.org
[5] http://www.tptp.org

Pattern	**#1**	**#2**	**#3**	**#4**	Total
hypernymy	2,559	458	–	–	**3,017**
cohyponymy	6,243	1,147	1,128	204	**8,722**
meronymy	6,669	1,852	1,556	393	**10,470**
attribute	3,092	600	685	109	**4,486**
event	12,575	3,220	3,357	721	**19,873**
Total	–	–	–	–	**46,568**

Table 3: CQs that results from BLESS

long faint constellation in the southern hemisphere near the equator stretching between Virgo and Cancer") are connected to $Clothing_c=$ and $AstronomicalBody_c+$ via *equivalence* and *instantiation*.

3 New WSD Proposal

In this section, we describe our proposal for the disambiguation of BLESS on the basis of the knowledge in SUMO and WordNet.

To this end, we conveniently use of our method for the evaluation of SUMO-based ontologies [10]. This method is an adaptation of the methodology for the design and evaluation of ontologies introduced in [17], which is based on the use of *competency questions* (CQs): problems that an ontology is expected to answer. The creation of CQs can be automatized by the use of few manually created *question patterns* (QPs), which exploit WordNet and its mapping into SUMO. Additionally, the evaluation of CQs can also be automatized by performing two dual tests per CQ FOL *automated theorem provers* (ATPs) such as Vampire: the first test is to check whether, as expected, the conjecture stated by the CQ is entailed by the ontology (*truth-test*); the second one is to check its complementary (*falsity-test*). If ATPs find a proof for either the truth- or the falsity-test, then the CQ is classified as *solved* (or *resolved*). In particular, the CQ is *passing*/*non-passing* if ATPs find a proof for the truth-test/falsity-test. Otherwise (that is, if no proof is found), the CQ is classified as *unresolved* or *unknown*.[6]

By using the above described method, our WSD proposal consists in choosing the most likely synset among the ones to which a given target belongs. For this purpose, we consider all the synsets to which words belong —both targets and their related words— and apply several predefined QPs to the resulting synset pairs. By using FOL ATPs, the most likely synset is decided to be the one with the largest difference between the amount of passing and non-passing CQs.

Next, we describe the process of creating CQs by means of the BLESS pair "*turtles* have *legs*", where *turtle* is the target that is related to *leg* by *meronymy*. In WordNet, *turtle* belongs to 2 synsets, while *leg* belongs to 9 different synsets. Therefore, we get 18 synset pairs for this BLESS pair. Among them, we now consider the following two ones:

1. $turtle^2_n$ and leg^1_n.
2. $turtle^1_n$ and leg^3_n

Regarding the first synset pair, $turtle^2_n$ ("*any of various aquatic and land reptiles having a bony shell and flipper-like limbs for swimming*") is connected to $Reptile_c+$ and leg^1_n ("*a structure in*

[6] Given a consistent ontology, ATPs cannot find a proof for both the truth- and the falsity-test.

Gold standard \ WordNet	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Total**
1	48	61	36	9	10	5	3	3	0	1	1	**177**
2	0	6	4	3	4	1	1	0	1	0	0	**20**
3	0	0	0	1	0	1	0	0	1	0	0	**3**
Total	**48**	**67**	**40**	**13**	**14**	**7**	**4**	**3**	**2**	**1**	**1**	**200**

Table 4: Distribution of BLESS targets in our gold standard and WordNet

animals that is similar to a human leg and used for locomotion") is connected to $Limb_c+$. Thus, by applying the first QP described in [4], we obtain the following CQ:

```
(exists (?X ?Y)
        (and
                (instance ?X Reptile)
                (instance ?Y Limb)
                (properPart ?Y ?X)))
```

With respect to the second synset pair, $turtle^1_n$ ("*a sweater or jersey with a high close-fitting collar*" is connected to $Clothing_c+$ and leg^3_n is connected to "*one of the supports for a piece of furniture*". Therefore, by applying the same QP we obtain

```
(exists (?X ?Y)
        (and
                (instance ?X Clothing)
                (instance ?Y Artifact)
                (properPart ?Y ?X)))
```

In BLESS, there are 14,400 word pairs related by *hypernymy*, *cohyponymy*, *meronymy*, *attribute* and *event* that yield 386,891 synset pairs. For each pair, we proceed as above by applying the corresponding QP according to the semantic relation that is used in the pair and the mapping relations that are used to connect the involved synsets to SUMO. In total, we consider 18 different QPs: 2 QPs for *hypernymy* and 4 QPs for each of the remaining relations. As result of this proces, we obtain 46,568 different CQs that are distributed as described in Table 3.

4 Experimentation

In this section, we report on the results that we have obtained at the disambiguation of the 200 targets in BLESS. For this purpose, in the next subsection we describe our evaluation framework, including the resources and remaining disambiguation systems that we consider for the evaluation of our proposal. Then, we summarize our experimentation results in Subsection 4.2. Finally, we discuss and analyze some disambiguation examples in Subsection 4.3.

4.1 Evaluation Framework

In order to evaluate our results, we have developed a *gold standard* for BLESS. For this purpose, a human expert has manually disambiguated the 200 targets of BLESS by taking into account

its context words. Since disambiguating BLESS is a difficult task even for humans, only 177 targets have been disambiguated to a single synset, while 20 targets have been disambiguated to 2 synsets and 3 targets to 3 synsets. In Table 4, we report the amount of targets that belongs to a range from 1 to 11 synsets in WordNet (in columns) and that are disambiguated to a range from 1 to 3 synsets in our gold standard (rows). For example, the BLESS target *dress* belongs to the context of *clothings* and appears in three WordNet synsets:

- $dress^1_n/frock^2_n$: "*a one-piece garment for a woman; has skirt and bodice*", connected to $Clothing_c+$.
- $attire^1_n/garb^1_n/dress^2_n$: "*clothing of a distinctive style or for a particular occasion*", connected to $Clothing_c+$.
- $apparel^1_n/wearing\_apparel^1_n/dress^3_n/clothes^1_n$: "*clothing in general*", connected to $Clothing_c+$.

And the BLESS target *villa* belongs to the context of *buildings* and also appears in three WordNet synsets:

- $Villa^1_n/Pancho\_Villa^1_n/Francisco\_Villa^1_n/Doroteo\_Arango^1_n$: "*Mexican revolutionary leader (1877-1923)*", connected to $Man_c@$.
- $villa^2_n$: "*detached or semidetached suburban house*", connected to $Building_c+$.
- $villa^3_n$: "*country house in ancient Rome consisting of residential quarters and farm buildings around a courtyard*", connected to $Building_c+$.
- $villa^4_n$: "*pretentious and luxurious country residence with extensive grounds*", connected to $StationaryArtifact_c+$.

The human expert undoubtedly disambiguates *dress* as $dress^1_n/frock^2_n$, while the target *villa* may refer to either $villa^2_n$, $villa^3_n$ or $villa^4_n$. Thus, the human expert selects the last three synsets (that is, $villa^2_n$, $villa^3_n$ and $villa^4_n$) as gold standard for *villa*.

Additionally, for a better evaluation of our proposal, we consider the results obtained by two different disambiguation systems: First, we take as baseline a random disambiguation system. The performance of this baseline system is calculated as the average ratio between the number of synsets related to a word according to our gold standard and the total number of synsets to which that word belongs in WordNet. Using the reported target distribution in Table 4, the resulting average ratio is 0.5538, from which we can conclude that the baseline performance is 55.35 %. Second, we have disambiguated BLESS using UKB [1], which is a well-known state-of-the-art tool in WSD, and compare the results with our gold standard. Totally, UKB correctly disambiguates 180 from 200 targets, thus its performance is 90.00 %.

4.2 Experimentation Results

Our experimentation has been performed by using Vampire v4.2.2 —which is the *CADE ATP System Competition* (CASC) FOF[7] division winner in 2017 [23, 27] and the latest available release[8] of Vampire at the time of our experimentation— in a Intel® Xeon® CPU E5-2640v3@2.60GHz with 2GB of RAM memory per processor. For each test, we have set an execution-time limit of 300 seconds and a memory limit of 2GB.[9] Totally, the experimentation

[7] First-Order Form non-propositional theorems (axioms with a provable conjecture).
[8] https://vprover.github.io/
[9] Parameters: `--proof tptp --output_axiom_names on --mode casc -t 300 -m 2048`

Pattern	#1		#2		#3		#4		Solved	
hypernymy	+1,027	−944	+173	−157	—	—	—	—	**1,201**	**39.81%**
cohyponymy	+712	−364	+189	−23	+216	−14	+27	−4	**1,549**	**17.76%**
meronymy	+785	−6	+37	−0	+109	−0	+0	−0	**937**	**8.95%**
attribute	+90	−84	+8	−1	+4	−3	+0	−0	**190**	**4.24%**
event	+0	−0	+0	−0	+0	−0	+0	−0	**0**	**0.00%**
Total	—	—	—	—	—	—	—	—	**3,877**	**8.33%**

Table 5: Summary of Experimentation Results

has required more than 320 days/processor of computation effort: 46,568 CQs, 2 tests per CQ and 300 seconds per test. All the required knowledge resources —the ontology Adimen-SUMO under CWA, the set of CQs and conjectures, the mapping between SUMO and WordNet v3.0, WordNet v3.0 relation pairs— and the resulting execution reports are available at https://adimen.si.ehu.es/web/AdimenSUMO.

We summarize our experimental results in Table 5, where CQs are organized by QP as in Table 3. For each QP, we provide the number of CQ classified as *passing* (prefixed by +) and *non-passing* (prefixed by −). Further, in the last two columns we provide the number and percentage of solved CQs.

Totally, only 8.33% of CQs are solved. However, this result is not surprising since we are exhaustively combining all the synsets in which a pair of words occur and, in fact, most of the synsets are not related. Further, the fact that all the CQs resulting from the QP *event* remain unsolved is also not surprising since the lack of knowledge about events in the ontology was also detected in the experimentation reported in [9] and confirmed in [6]. On the contrary, the large amount of solved CQs belong to the QPs *hypernymy* and *cohyponymy* as expected, because the best results in the experimentations reported in [9, 10] are obtained for the CQs based on *hypernymy*/*hyponymy*.

On the basis of these results, then we disambiguate BLESS by choosing for each target the synset with the largest difference between the amount of passing and non-passing CQs. By proceeding in this way, our proposal can correctly disambiguate 140 targets, thus obtaining a performance of 70.00 %. Consequently, our proposal clearly outperforms the baseline (55.35 %) although is still far from the performance of UKB (90.00 %).

By a deeper analysis of the disambiguation results, we check that UKB disambiguates 39 targets that are not correctly disambiguated using our proposal. On the contrary, our proposal is able to disambiguate 14 targets that UKB cannot do. Therefore, the upper bound performance of a disambiguation system that combines UKB and our proposal increases up to 97.00 %, which opens new research lines for the improvement of state-of-the-art disambiguation systems.

4.3 Analysis and Discussion

In this section, we introduce some examples of BLESS targets that are not correctly disambiguated by UKB or our proposal and try to detect the causes of disambiguation failures.

The first example is the target *donkey* that belongs to the context of *ground mammals*. This target is related in BLESS with the following words:

- By *hypernymy*: *animal*, *mammal*, *vertebrate*, ...
- By *cohyponymy*: *fox*, *lion*, *pig*, ...

- By *meronymy*: *neck*, *ear*, *leg*, ...
- By *attribute*: *big*, *large*, *stubborn*, ...
- By *event*: *die*, *eat*, *live*, ...

donkey is correctly disambiguated by UKB to $domestic\_ass^1_n/donkey^2_n/Equus\_asinus^2_n$ ("*domestic beast of burden descended from the African wild ass; patient but stubborn*"), which is connected to $HoofedMammal_c+$. On the contrary, our proposal incorrectly disambiguates *donkey* to $donkey^1_n$ ("*the symbol of the Democratic Party; introduced in cartoons by Thomas Nast in 1874*"), which is connected to $Icon_c+$. We have identified three possible causes of this failure. First, the SUMO concept $HoofedMammal_c$, to which the target is related, is under-axiomatized in the ontology and, thus, the required CQs are not entailed by the ontology. Second, the mapping of the related words is not suitable and, consequently, the resulting CQs and the context of the target do not semantically match. Third, our proposal can be improved by conveniently weighting the amount of passing and non-passing CQs. This way, we can do additional experiments in order to find the weight values that returns the optimal disambiguation results.

The second example is the target *fighter* that belongs to the context of *vehicles*. This target is related in BLESS with the following words:

- By *hypernymy*: *plane*, *vehicle*, *transport*, ...
- By *cohyponymy*: *car*, *bus*, *train*, ...
- By *meronymy*: *missile*, *seat*, *metal*, ...
- By *attribute*: *big*, *destructive*, *lethal*, ...
- By *event*: *leave*, *go*, *run*, ...

The word *fighter* occurs in 3 WordNet synsets:

- $combatant^1_n/battler^1_n/belligerent^1_n/fighter^1_n/crapper^1_n$ ("*someone who fights (or is fighting)*"), which is connected to $SocialRole_c+$.
- $fighter^2_n/fighter\_aircraft^1_n/attack\_aircraft^1_n$ ("*a high-speed military or naval airplane designed to destroy enemy aircraft in the air*"), which is connected to $AirCraft_c+$.
- $champion^2_n/fighter^3_n/hero^3_n/paladin^1_n$ ("*someone who fights for a cause*"), which is connected to $Human_c+$.

Our proposal correctly disambiguates the target *fighter* to the second synset (that is, $fighter^2_n/fighter\_aircraft^1_n/attack\_aircraft^1_n$) while UKB incorrectly chooses the third one: that is, $champion^2_n/fighter^3_n/hero^3_n/paladin^1_n$. In this case, we think that the cause of the problem is that this sense of *fighter* is not so common in corpora.

The last example is given by the target *herring*, which belongs to the context of *water animals*. The word *herring* is related in BLESS with:

- By *hypernymy*: *fish*, *food*, *animal*, ...
- By *cohyponymy*: *salmon*, *tuna*, *cod*, ...
- By *meronymy*: *eye*, *skin*, *tail*, ...

- By *attribute*: *edible*, *fresh*, *small*, ...
- By *event*: *cook*, *live*, *eat*, ...

In our gold standard, *herring* is disambiguated to the synset $herring^2_n$/$Clupea\_harangus^1_n$ ("*commercially important food fish of northern waters of both Atlantic and Pacific*"), which is connected to $Fish_c+$. However, both our proposal and UKB incorrectly disambiguate the target *herring* to $herring^1_n$ ("*valuable flesh of fatty fish from shallow waters of northern Atlantic or Pacific; usually salted or pickled*"), which is connected to $Meat_c+$. We think that the cause of this failure is that some contexts in BLESS are ambiguous or not so fine grained as in WordNet. In the case of *herring*, we have that its both meanings (animal and flesh) are mixed in BLESS: the target is related with both *fish* and *food* by *hypernymy*. Further, the role in the relation of *herring* with the verb *eat* by *event* may be *agent* (to eat) or *patient* (to be eaten). Similarly, our proposal also suffers from ambiguity when several synsets are connected to the same SUMO concepts and, consequently, there is no possible disambiguation.

5 Conclusions and Future Work

In this paper, we have demonstrated that, although the ontology and its mapping to WordNet can be further improved, Adimen-SUMO can be applied to NLP tasks in its current state with the help of Vampire v4.2.2., in particular, for WSD.

In the future, we plan to continue improving the knowledge resources and tools in order to obtain better results. Among others, we consider two main areas for this purpose.

On one hand, we are improving the involved knowledge resources in several ways. First, we are correcting mapping and knowledge errors by applying both black- [10] and white-box testing [7] techniques. Second, we are improving the quality of the mapping between WordNet and SUMO by focusing on particular relations such as *meronymy* [4, 5] and metonymy [16]. Third, we are improving the knowledge in Adimen-SUMO by providing stronger axiomatizations and definitions of the concepts in the ontology. In particular, we are moving most of the restrictions from the level of objects to classes, we are applying the CWA [3] and plan to apply other assumptions such as *Unique Name Assumption* (UNA) [24], and we are implementing some optimizations on the resulting FOL formula such as removing non-constant function symbols. Further, we are also considering the possibility of preprocessing the formula in order to help the work of ATPs.

On the other hand, we want to improve the tools that are used to work with Adimen-SUMO. More concretely, we are trying other ATP systems like E [25], CVC4 [12] or iProver [18], and plan to improve the application of Vampire v4.2.2. by searching for specialized portfolios and by increasing the available resource limits. Additionally, we are developing *ad hoc* reasoning tools for the particular case of commonsense knowledge.

Acknowledgments

This work has been partially funded by the the project DeepReading (RTI2018-096846-B-C21) supported by the Ministry of Science, Innovation and Universities of the Spanish Government, and GRAMM (TIN2017-86727-C2-2-R) supported by the Ministry of Economy, Industry and Competitiveness of the Spanish Government, the Basque Project LoRea (GIU18/182) and BigKnowledge – *Ayudas Fundación BBVA a Equipos de Investigación Científica 2018*.

References

[1] E. Agirre, O. López de Lacalle, and A. Soroa. The risk of sub-optimal use of open source NLP software: UKB is inadvertently state-of-the-art in knowledge-based WSD. *CoRR*, abs/1805.04277, 2018.

[2] E. Agirre and P. Edmonds. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media, 2007.

[3] J. Álvez, I. Gonzalez-Dios, and G. Rigau. Applying the Closed World Assumption to SUMO-based ontologies. *CoRR*, abs/1808.04620, 2018.

[4] J. Álvez, I. Gonzalez-Dios, and G. Rigau. Cross-checking WordNet and SUMO using meronymy. In *Proc. of the 11th Int. Conf. on Language Resources and Evaluation (LREC 2018)*, 2018.

[5] J. Álvez, I. Gonzalez-Dios, and G. Rigau. Validating WordNet meronymy relations using Adimen-SUMO. *CoRR*, abs/1805.07824, may 2018.

[6] J. Álvez, I. Gonzalez-Dios, and G. Rigau. Commonsense reasoning using WordNet and SUMO: a detailed analysis. In P. Vossen and C. Fellbaum, editors, *Proc. of the 10th Global WordNet Conference (GWC 2019)*, 2019.

[7] J. Álvez, M. Hermo, P. Lucio, and G. Rigau. Automatic white-box testing of first-order logic ontologies. *Journal of Logic and Computation*, 29(5):723–751, 2 2019.

[8] J. Álvez, P. Lucio, and G. Rigau. Adimen-SUMO: Reengineering an ontology for first-order reasoning. *Int. J. Semantic Web Inf. Syst.*, 8(4):80–116, 2012.

[9] J. Álvez, P. Lucio, and G. Rigau. Black-box testing of first-order logic ontologies using WordNet. *CoRR*, abs/1705.10217, 2017.

[10] J. Álvez, P. Lucio, and G. Rigau. A framework for the evaluation of SUMO-based ontologies using WordNet. *IEEE Access*, pages 1–20, 2019.

[11] M. Baroni and A. Lenci. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics, 2011.

[12] M. Deters, A. Reynolds, T. King, C. W. Barrett, and C. Tinelli. A tour of CVC4: how it works, and how to use it. In *Proc. of the Formal Methods in Computer-Aided Design (FMCAD 2014)*, page 7, 2014.

[13] N. Drummond and R. Shearer. The open world assumption. In *eSI Workshop: The Closed World of Databases meets the Open World of the Semantic Web*, volume 15, 2006.

[14] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[15] M. R. Genesereth, R. E. Fikes, D. Brobow, R. Brachman, T. Gruber, P. Hayes, R. Letsinger, V. Lifschitz, R. Macgregor, J. Mccarthy, P. Norvig, R. Patil, and L. Schubert. Knowledge Interchange Format version 3.0 reference manual. Technical Report Logic-92-1, Stanford University, Computer Science Department, Logic Group, 1992.

[16] I. Gonzalez-Dios, J. Álvez, and G. Rigau. Exploiting metonymy from available knowledge resources. In *Proc. of the 20th Int. Conf. on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*, volume (to appear) of *LNCS*. Springer, 2019.

[17] M. Grüninger and M. S. Fox. Methodology for the design and evaluation of ontologies. In *Proc. of the Workshop on Basic Ontological Issues in Knowledge Sharing (IJCAI 1995)*, 1995.

[18] K. Korovin. Inst-Gen - A modular approach to instantiation-based automated reasoning. In *Programming Logics - Essays in Memory of Harald Ganzinger*, pages 239–270, 2013.

[19] L. Kovács and A. Voronkov. First-order theorem proving and Vampire. In N. Sharygina and H. Veith, editors, *Computer Aided Verification*, LNCS 8044, pages 1–35. Springer, 2013.

[20] I. Niles and A. Pease. Towards a standard upper ontology. In Guarino N. et al., editor, *Proc. of the 2nd Int. Conf. on Formal Ontology in Information Systems (FOIS 2001)*, pages 2–9. ACM, 2001.

[21] I. Niles and A. Pease. Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In H. R. Arabnia, editor, *Proc. of the IEEE Int. Conf. on Inf. and Knowledge Engin. (IKE 2003)*, volume 2, pages 412–416. CSREA Press, 2003.

[22] A. Pease. Standard Upper Ontology Knowledge Interchange Format. Retrieved June 18, 2009, from http://sigmakee.cvs.sourceforge.net/sigmakee/sigma/suo-kif.pdf, 2009.

[23] F.J. Pelletier, G. Sutcliffe, and C.B. Suttner. The Development of CASC. *AI Communications*, 15(2-3):79–90, 2002.

[24] S. J. Russell and P. Norvig. *Artificial Intelligence - A Modern Approach (3. internat. ed.)*. Pearson Education, 2010.

[25] S. Schulz. E - A brainiac theorem prover. *AI Communications*, 15(2-3):111–126, 2002.

[26] G. Sutcliffe. The TPTP problem library and associated infrastructure. *J. Automated Reasoning*, 43(4):337–362, 2009.

[27] G. Sutcliffe and C. Suttner. The State of CASC. *AI Communications*, 19(1):35–48, 2006.