SemR-11: A Multi-Lingual Gold-Standard for Semantic Similarity and Relatedness for Eleven Languages

Siamak Barzegar[1], Brian Davis[2], Manel Zarrouk[1] Siegfried Handschuh[3], Andre Freitas[4]
[1]Insight Centre for Data Analytics, National University of Ireland, Galway
siamak.barzegar,manel.zarrouk@insight-centre.org
[2]Department of Computer Science, Maynooth University
brian.davis@mu.ie
[3]Department of Mathematics and Computer Science, University of Passau
siegfried.handschuh@uni-passau.de
[4]School of Computer Science, The University of Manchester
andre.freitas@manchester.ac.uk

Abstract

This work describes SemR-11, a multi-lingual dataset for evaluating semantic similarity and relatedness for 11 languages (German, French, Russian, Italian, Dutch, Chinese, Portuguese, Swedish, Spanish, Arabic and Persian). Semantic similarity and relatedness gold standards have been initially used to support the evaluation of semantic distance measures in the context of linguistic and knowledge resources and distributional semantic models. SemR-11 builds upon the English gold-standards of Miller & Charles (MC), Rubenstein & Goodenough (RG), WordSimilarity 353 (WS-353), and Simlex-999, providing a canonical translation for them. The final dataset consists of 15,917 word pairs and can be used to support the construction and evaluation of semantic similarity/relatedness and distributional semantic models. As a case study, the SemR-11 test collections was used to investigate how different distributional semantic models built from corpora in different languages and with different sizes perform in computing semantic relatedness similarity and relatedness tasks.

Keywords: Gold standard, Semantic Similarity, Semantic Relatedness, Multi-linguality, Word-embeddings

1. Motivation

The ability to automatically determine and quantify the degree of semantic similarity and semantic relatedness between pairs of words or expressions is one of the archetypal tasks for assessing the ability of a system to perform semantic interpretation. The ability to quantify semantic relatedness can provide a lightweight semantic interpretation operation which can be applied in different areas of Artificial Intelligence, Natural Language Processing and Information Retrieval. Examples of applications include coping with lexical and semantic gaps in Question Answering Systems (Freitas, 2015; Freitas and Curry, 2014), using the semantic relatedness score as a ranking function in Information Retrieval systems (Freitas et al., 2012) and serving as a semantic scoping mechanism in deductive/abductive methods (Freitas et al., 2014).

Due to its simplicity in comparison to other tasks such as Question Answering, Text Entailment and Machine Translation, semantic similarity and relatedness gold standards have been initially used to support the evaluation of the interaction between semantic distance measures and of linguistic and knowledge resources (Resnik, 1995; Lin, 1991; Wu and Palmer, 1994; Agirre et al., 2009). As the conditions to process large-scale corpora emerged, distributional semantic models automatically built from textual corpora were created (Turney and Pantel, 2010a) using, in most cases, a vector space representation of meaning. As distributional semantic models can induce modes with a more comprehensive underlying vocabulary and also capture a broader set of semantic relations, new gold-standards emerged (Finkelstein et al., 2001), evolving from capturing semantic similarity to semantic relatedness behavior. More recently, the creation of neural/predictive word embedding models (Mikolov et al., 2013; Pennington et al., 2014) pushed semantic similarity and relatedness gold-standards to evolve in the direction of quantifying more fine-grained semantic relations (Hill et al., 2015).

Currently, most of the existing gold-standards for evaluating semantic similarity and relatedness have focused on the English language, with some initiatives providing initial gold-standards for few other languages (Faruqui and Dyer, 2014). This paper describes SemR-11, a multi-lingual gold-standard which aims at generalizing existing semantic similarity and relatedness gold-standards to 11 languages (German, French, Russian, Italian, Dutch, Chinese, Portuguese, Swedish, Spanish, Arabic and Persian). The resource is built using a principled translation method over four reference gold-standards: Miller & Charles (Miller and Charles, 1991), Rubenstein & Goodenough (Rubenstein and Goodenough, 1965), WS-353 (Finkelstein et al., 2001) and Simlex-999 (Leviant and Reichart, 2015). The final resource contains in total 15,917 word pairs.

The resource aims to contribute to research in the following directions:

- Supporting the development of linguistic resources and distributional semantic models for non-English languages.
- Providing a comparative framework for analyzing the impact of language structural features and types (e.g. analytic, isolating and synthetic languages) in the development of semantic relatedness models.

- Evaluating the use of machine translation to support semantic similarity and relatedness (Freitas et al., 2016).
- Creating semantic similarity and relatedness models which work on languages not having a high-volume supporting corpora.

This paper is organized as follows: Section 2. describes the state-of-the-art in existing gold-standards for semantic similarity and relatedness computations as well as their language variants; Section 3. describes the English gold-standards which were used as a reference for the machine translation process; Section 4. describes the SemR-11 gold-standard and its creation process.

2. Related Work

Camacho-Collados et al. (2017) developed a multi-lingual gold-standard which includes 518 word pairs for five languages (English, German, Italian, Spanish and Persian). It is composed of nominal pairs of multi-word expressions, domain-specific terms and named entities that are manually scored between 0 to 4 where 0 indicates that they are completely dissimilar and 4 denotes that the two words are synonymous. This dataset (Camacho-Collados et al., 2017) focuses on semantic similarity.
Bruni et al. (2014) introduced a test collection containing 3000 word pairs. The MEN dataset obtained by crowdsourcing using Amazon Mechanical Turk [1] via the CrowdFlower[2] interface. The dataset focuses on semantic relatedness pairs on the English language (similarly to the WS-353 dataset (Finkelstein et al., 2001)). They developed it, specifically to test multimodal models. Compared to WS-353, MEN is sufficiently large, and the human judgments are relative rather than absolute. At (Bruni et al., 2014), each rater chose the word pair that was more similar out of two random pairs of words. They used this technique to have a comparative judgment rather than absolute scores for single pairs, which was used in the WS-353.
Agirre et al. (2009) split the WS-353 (Finkelstein et al., 2001) into two test collections (WS-Sim and WS-Rel) containing 203 and 252 word pairs on the English language, respectively. *WS-Sim* focused on only measuring similarity, and the other one on only relatedness.

3. Reference Gold-standards

SemR-11 consists of the translation of four semantic similarity and relatedness gold-standards: Miller & Charles (MC) (Miller and Charles, 1991), Rubenstein & Goodenough (RG) (Rubenstein and Goodenough, 1965), WordSimilarity 353 (WS-353) (Finkelstein et al., 2001) and Simlex-999 (Leviant and Reichart, 2015). These four datasets were selected for being consensual gold-standards for the evaluation of semantic similarity and relatedness models.
The problem of measuring the semantic similarity and relatedness of two concepts can be stated as follows: given two concepts A and B, determine a numerical measure f(A, B) which expresses the semantic similarity or relatedness between concepts A and B. The notion of semantic similarity is associated with taxonomic (is-a) relations, while semantic relatedness represents more general relations. *Car* and *train* are examples of similar concepts (both share a common taxonomic ancestor, *vehicle*) while *car* and *wheel* are related concepts (a wheel is part of a car). As a consequence, semantic similarity is considered a particular case of semantic relatedness. Alternatively semantic similarity can also be defined as two concepts sharing a high number of salient features (attributes): synonymy (car/automobile), hyperonymy (car/vehicle), co-hyponymy (car/van/truck), while semantic relatedness can be defined as two words semantically associated without being necessarily similar: function (car/drive), meronymy (car/tyre), location (car/road), attribute (car/fast) (Freitas, 2015).
The gold standards are described below:

Pairs	Simlex-999	WS-353
closet - clothes	1.15	8.0

Table 1: Semantic Similarity vs Semantic Relatedness

- **Wordsimilarity 353:** WS-353 (Finkelstein et al., 2001) is certainly the most popular evaluation gold standard for distributional semantic models. The dataset is focused on semantic relatedness. The dataset contains two subsets: *set 1* (153 word pairs, evaluated by 13 subjects), and *set 2* (200 word pairs evaluated by 16 subjects) each one containing pairs from different parts-of-speech, a proper noun and pairs involving subjective bias.
- **Rubenstein & Goodenough:** RG (Rubenstein and Goodenough, 1965) contains 65 pairs which are often used to evaluate Distributional Semantic Models. RG reflects similarity of words rather than their relatedness. It is build by rating of 15 annotators to score the semantic similarity of each pair.
- **Miller & Charles:** MC (Miller and Charles, 1991) is a subset of 30 noun pairs from the RG gold standard which are re-annotated following new similarity guidelines. Ten pairs were selected from the highest level (between 3 and 4 on a scale from 0 to 4), ten pairs from the intermediate level (between 1 and 3), and ten pairs from the lowest level (0 to 1) of semantic similarity.
- **SIMLEX-999:** Simlex-999 (Hill et al., 2016; Leviant and Reichart, 2015) is aimed to measure how well Distributional Semantic Models capture similarity, rather than relatedness. Simlex-999 contains a range of 111 adjective, 666 noun and 222 verb pairs with an independent rating for each pair. It was built by using 500 annotators via Amazon Mechanical Turk.

4. SemR-11

The process of creating SemR-11 (Table 3) consisted in the translation of the three gold-standards WS-353, MC,

[1] https://www.mturk.com/mturk/welcome
[2] http://crowdflower.com/

Language	Parametes	MC	RG	WS-353	SIMLEX-999
German	# of Tokens	40	52	431	1094
	Vocabulary Size	40	52	431	1094
French	# of Tokens	37	45	430	1106
	Vocabulary Size	37	43	424	1097
Russian	# of Tokens	38	48	435	-
	Vocabulary Size	36	46	426	-
Italian	# of Tokens	34	43	426	1051
	Vocabulary Size	34	43	424	1051
Dutch	# of Tokens	37	45	426	1025
	Vocabulary Size	37	45	426	1018
Chinese	# of Tokens	37	51	471	-
	Vocabulary Size	37	51	471	-
Portuguese	# of Tokens	37	46	434	1149
	Vocabulary Size	37	46	434	1141
Swedish	# of Tokens	35	44	430	1002
	Vocabulary Size	35	44	430	995
Spanish	# of Tokens	35	44	437	993
	Vocabulary Size	35	44	437	991
Arabic	# of Tokens	38	54	448	-
	Vocabulary Size	36	49	448	-
Persian	# of Tokens	34	43	456	-
	Vocabulary Size	34	43	436	-

Table 2: The vocabulary and token distribution for each language of four gold-standards

Language	SemR-11				SE17 T2
	MC	RG	WS 353	Simlex 999	
German	✓	✓	✓	✓	✓
French	✓	✓	✓	✓	
Russian	✓	✓	✓		
Italian	✓	✓	✓	✓	✓
Dutch	✓	✓	✓	✓	
Chinese	✓	✓	✓		
Portuguese	✓	✓	✓	✓	
Swedish	✓	✓	✓	✓	
Spanish	✓	✓	✓	✓	✓
Arabic	✓	✓	✓		
Persian	✓	✓	✓		✓

Table 3: SemR-11 and its relation to existing multi-lingual gold standards.

English	Portuguese
hard;difficult;9.69	difícil;difícil;10
apparent;obvious; 9.08	visível;óbvio;9.15
disease;infection;7.08	doença;infecção;4.46

Table 4: Comparison between English and Portuguese gold standards.

RG for eleven languages and of the Simlex-999 for seven European languages (German, French, Italian, Dutch, Portuguese, Swedish and Spanish). Also SemR-11 has been compared with an existing multi-lingual gold standards [3]. The word pairs were translated by paid professional translators[4], skilled in data localisation tasks.
All translated pairs followed the protocol below:

1. Given a pair of words, translators should assume the most similar senses associated with the pair.
2. Translators should preserve the lexical category of the sense identified for that word.

In the end, 15,917 word pairs were translated to 11 languages. Table 2 quantifies the vocabulary and token distribution for each language.
The datasets are available on the Web[5].
The SemR-11 gold-standard assumes that the translations are preserving the similarity and relatedness scores of their original English human annotation. The target task was described to the human translators, who had access to the word pairs and scores.

English	French
woman;wife;283	Femme;femme;10
girl;child;4.77	Fille;enfant;5
understand;know;5.69	Comprendre;connaître;6.92

Table 5: Comparison between English and French gold standards

[3]SemEval-2017 Task 2
[4]Lionbridge Natural Language Solutions
[5]https://github.com/Lambda-3/Gold-Standards/tree/master/SemR-11

word-Pairs	MC	RG	WS-353	Simlex-999
English	food;rooster	monk;oracle	closet;clothes	clothes;closet
German	nahrung;hahn	mönch;orakel	Wandschrank;Kleidung	Kleider;Schrank
French	nourriture;coq	moine;oracle	cabinet;vêtements	vêtements;placard
Russian	еда;петух	монах;оракул	стенной шкаф;одежда	-
Italian	cibo;gallo	monaco;oracolo	ripostiglio;vestiti	vestiti;armadio
Dutch	voedsel;haan	monnik;orakel	kast;kleren	kleding;kast
Chinese	食物;公鸡	僧侣;甲骨文	壁橱;衣服	-
Portuguese	comida;galo	monge;oráculo	armário;roupas	roupas;roupeiro
Swedish	mat;tupp	munk;orakel	garderob;kläder	kläder;förråd
Spanish	comida;gallo	monje;oráculo	armario;ropa	ropa;armario
Arabic	طعام;ديك	راهب;وحي	خزانة;ملابس	-
Persian	غذا;خروس	راهب;وحی	گنجه;لباس	-

Table 6: Examples with all the languages for each of four datasets

Tables 4 and 5 show examples of translated pairs of Simlex-999 test collection (with the associated average similarity score) into Portuguese and French languages, respectively, while Table 6 provides example of word-pairs for each language and dataset.

5. Use Case

Distributional Semantic Models (DSM) are consolidating themselves as fundamental components for supporting automatic semantic interpretation in different application scenarios in natural language processing. From *question answering systems*, to *semantic search* and *text entailment*, distributional semantic models support a scalable approach for representing the meaning of words, which can automatically capture comprehensive associative commonsense information by analysing word-context patterns in large-scale corpora in an unsupervised or semi-supervised fashion (Freitas, 2015; Turney and Pantel, 2010b; Sales et al., 2016).

The SemR-11 test collection was used by Freitas et al.(2016), Sales et al.(2018) and Barzegar et al.(2018) to evaluate how different distributional semantic models built from corpora in different languages and with different sizes, perform in computing semantic relatedness similarity and relatedness tasks. Additionally, SemR-11 was used to analyze the role of machine translation approaches to support the construction of high-quality distributional vectors and computing semantic similarity & relatedness measures for other languages.

6. Acknowledgments

This publication has emanated from research funded in part from the European Union's Horizon 2020 research and innovation programme under grant agreement No 645425 SSIX and Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

We would like in particular to thank Alexandros Poulis and Juha Vilhunen from the Global Services for Machine Intelligence Group, Lionbridge Finland [6] ensuring the production word of high quality translations for our similarity datasets.

7. Reference

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and 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*, pages 19–27. Association for Computational Linguistics.

Barzegar, S., Davis, B., Handschuh, S., and Freitas, A. (2018). Multilingual semantic relatedness using lightweight machine translation. In *IEEE International Conference on Semantic Computing*. IEEE.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(2014):1–47.

Camacho-Collados, J., Pilehvar, M. T., Collier, N., and Navigli, R. (2017). Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017). Vancouver, Canada*.

Faruqui, M. and Dyer, C. (2014). Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, USA, June. Association for Computational Linguistics.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Freitas, A. and Curry, E. (2014). Natural language queries over heterogeneous linked data graphs: A distributional-compositional semantics approach. In *Proceedings of*

[6] https://www.lionbridge.com/en-us/global-services-for-machine-intelligence

the 19th international conference on Intelligent User Interfaces, pages 279–288. ACM.

Freitas, A., Curry, E., and O'Riain, S. (2012). A distributional approach for terminological semantic search on the linked data web. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 384–391. ACM.

Freitas, A., da Silva, J. C. P., Curry, E., and Buitelaar, P. (2014). A distributional semantics approach for selective reasoning on commonsense graph knowledge bases. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, pages 21–32. Springer.

Freitas, A., Barzegar, S., Sales, J. E., Handschuh, S., and Davis, B. (2016). Semantic relatedness for all (languages): A comparative analysis of multilingual semantic relatedness using machine translation. In *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20*, pages 212–222. Springer.

Freitas, A. (2015). *Schema-agnositc queries over large-schema databases: a distributional semantics approach.* Ph.D. thesis, Digital Enterprise Research Institute (DERI), National University of Ireland, Galway.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Hill, F., Reichart, R., and Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Leviant, I. and Reichart, R. (2015). Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR Workshop Papers*.

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

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

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Sales, J. E., Freitas, A., Davis, B., and Handschuh, S. (2016). A compositional-distributional semantic model for searching complex entity categories. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 199–208.

Sales, J. E., Souza, L., Barzegar, S., Davis, B., Freitas, A., and Handschuh, S. (2018). Indra: A word embedding and semantic relatedness server. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Turney, P. D. and Pantel, P. (2010a). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Turney, P. D. and Pantel, P. (2010b). From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January.

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.