BabelEnconding at SemEval-2020 Task 3: Contextual Similarity as a Combination of Multilingualism and Language Models

Lucas R. C. Pessutto[1], Tiago de Melo[2], Viviane P. Moreira[1], Altigran da Silva[3]
[1]Institute of Informatics – UFRGS – Brazil; [2]Amazonas State University – UEA – Brazil;
[3]Federal University of Amazonas – UFAM – Brazil
lrcpessutto@inf.ufrgs.br, tmelo@uea.edu.br,
viviane@inf.ufrgs.br, alti@icomp.ufam.edu.br

Abstract

This paper describes the system submitted by our team (BabelEnconding) to SemEval-2020 Task 3: Predicting the Graded Effect of Context in Word Similarity. We propose an approach that relies on translation and multilingual language models in order to compute the contextual similarity between pairs of words. Our hypothesis is that evidence from additional languages can leverage the correlation with the human generated scores. BabelEnconding was applied to both subtasks and ranked among the top-3 in six out of eight task/language combinations and was the highest scoring system three times.

1 Introduction

Word similarity is a key task in Natural Language Processing (NLP) applications. Language models, such as word embeddings (Mikolov et al., 2013) create vector representations for the words that are able to capture syntactic and semantic relationships. These representations became very popular in the last few years as they have boosted the performance of several NLP tasks. However, since each word is represented by a fixed vector these techniques have problems dealing with polysemous words and identifying subtle meaning changes between different sentences. On the other hand, state-of-the-art language models, like BERT (Devlin et al., 2019) provide a contextualized word representation – the representation of a word relies on its context, which means that the same word may have different representations through the sentences. Thus, BERT models are more suitable for handling polysemous words.

Task 3 in SemEval 2020 – *Predicting the Graded Effect of Context in Word Similarity* (Armendariz et al., 2020a) was motivated by this improvement on language models. The task aims at the design of a similarity measure which captures the human perception of the meaning of words. For that purpose, task organizers built and annotated datasets in four languages – English, Croatian, Finnish, and Slovenian. Each entry in a dataset consists of two target words and two contexts, where each one is a piece of text containing both target words. The global task is divided into two subtasks: 1) predicting the change in the human annotator's scores of similarity when presented with the same pair of words within two different contexts; and 2) predicting the human scores of similarity for a pair of words within two different contexts.

In this paper, we describe BabelEnconding, an approach that relies on machine translation and multilingual language models to evaluate the contextual similarity of pairs of words. Our hypothesis is that having similarity information from more languages helps decide on how similar the words are.

Considering the eight combinations of language/subtask, BabelEnconding was ranked among the top-3 competitors six times, and was the top scoring method in three cases. Our additional experiments in English and Croatian showed that adding more languages noticeably improved the results for Croatian in both subtasks. In English, the gain was small and happened only in Subtask 2.

2 Background and Related Work

The Distributional Hypothesis (Harris, 1954) states that the meaning of a word changes depending on the context it is used. At the same time, this hypothesis also states that if two words tend to be used in the same

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 59–66
Barcelona, Spain (Online), December 12, 2020.

contexts, then they are likely to be more similar. This claim inspired many solutions in NLP that based solely on the distribution of words in the corpus (Fernández et al., 2016; Wang et al., 2020; Lüddecke et al., 2019). Word embeddings and language models, for example, are among these solutions. The idea is to represent words in a vector space in such a way that the semantic similarity between words is preserved. In the past few years, techniques to build language models became very popular. Word2vec (Mikolov et al., 2013) is an efficient and fast training method for word embeddings, based on co-occurrence statistics. The authors devised two model architectures for the word vectors training – continuous bag of words and skip-grams. Both approaches consist of neural networks trained to predict neighbor contextual words. Despite its ability in mapping linguistic regularities present in documents, this language model produces a unique representation for each word in the vocabulary, which prevents the differentiation of word senses.

In state-of-the-art language models, such as BERT (Devlin et al., 2019), the context of a word is taken into account in its representation. These models are trained over a large corpus to predict missing tokens which are removed from the original sentences. An advantage of BERT over Word2vec is that it creates different representations for the same word depending on the context in which the word appears. Another advantage of BERT-like models is that they can be specialized for a specific task with few training epochs.

Solutions for measuring contextual similarity between word pairs and word-sense disambiguation benefited from BERT-like language models. Enriched models were designed (Levine et al., 2019; Peters et al., 2019; Scarlini et al., 2020), and new datasets such as the Word-in-Context Dataset (Pilehvar and Camacho-Collados, 2018) and CoSimLex (Armendariz et al., 2020b) were assembled. Word-sense disambiguation can also take advantage of multilingualism. Some works have employed parallel/comparable corpora (Banea and Mihalcea, 2011; Dandala et al., 2013) and translation (Carpuat, 2013) to that task. Multilingual resources, such as Multi-SimLex (Vulic et al., 2020), were also developed and yielded improvements compared with the monolingual version.

3 BabelEnconding

Our proposed solution, called BabelEnconding, works in two phases and its overall process is depicted in Figure 1. The input is a pair of words and two sentences (contexts) containing both words of interest. More formally, let $S_1 = \{w_1^1, w_2^1, \ldots, w_i^1\}$ and $S_2 = \{w_1^2, w_2^2, \ldots, w_j^2\}$ be two sentences, where there is a pair of words $p = \langle w_a, w_b \rangle \in S_1$ and S_2. For example, let S_1 = *"Her prison* ***cell*** *was almost an improvement over her* ***room*** *at the last hostel"* and S_2 = *"His job didn't leave much* ***room*** *for a personal life. He knew much more about human* ***cells*** *than about human feelings"* be two sentences, where the pair of words $p = \langle room, cell \rangle$.

In the first phase of BabelEnconding, both input sentences S_1 and S_2 are translated into a set of k languages $L = \{l_1, l_2, \ldots, l_k\}$. This process will produce a set of translated sentences $S^{l_i} = \left\{S_1^{l_i}, S_2^{l_i}\right\}$, which corresponds to the translation of the original sentences, into each language $l_i \in L$. Then, the words of interest are identified in the translated text, generating two sets $p_{s_1}^{l_i}$ and $p_{s_2}^{l_i}$. In this example,

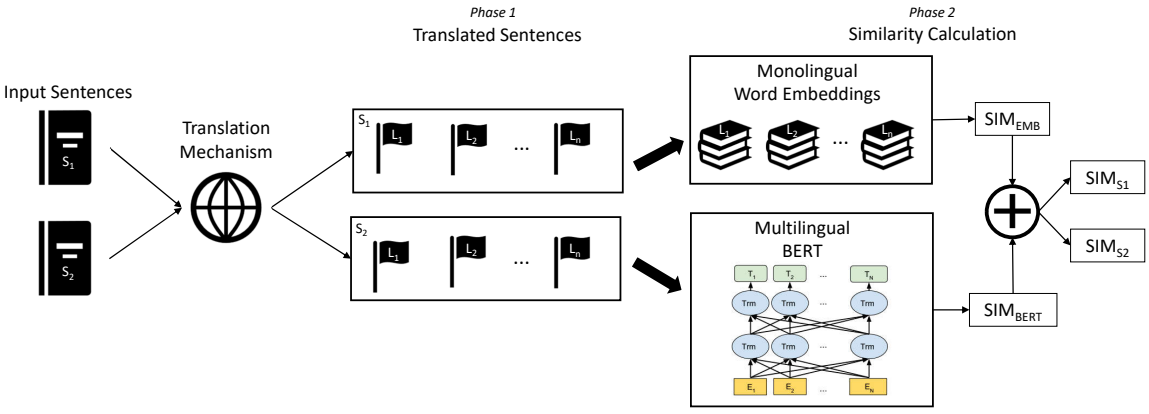


Figure 1: Overview of BabelEnconding

considering $L = \{\text{Italian}, \text{Portuguese}\}$, the pairs of words of interest are translated as $p_{s_1}^{\text{IT}} = \langle \text{cella}, \text{stanza} \rangle$, $p_{s_1}^{\text{PT}} = \langle \text{cela}, \text{quarto} \rangle$ from S_1 and $p_{s_2}^{\text{IT}} = \langle \text{spazio}, \text{cellule} \rangle$, $p_{s_2}^{\text{PT}} = \langle \text{espaço}, \text{células} \rangle$ from S_2.

In the second phase, with the translated sentences, we evaluate the similarity between the pair of words in p for each language in L separately in two ways: (i) using word embeddings and (ii) BERT. Finally, BabelEnconding calculates a weighted average between word embeddings and BERT similarities. These similarities are used to address both subtasks.

Word Embedding Similarity consists in taking the cosine similarity between the word vectors of the two words in each language. We rely on pre-trained monolingual word embeddings to represent the words. The context is not used in this similarity measure since there is a fixed vector for each word.

BERT Similarity requires inferring the word embedding representation of words in BERT models, taking context into consideration. The context is the sentence (S_m) containing the two words. This process was done summing the last four hidden layers of the BERT model. This choice was made based on the good results achieved by Devlin et al. (2019) in the Named Entity Recognition task.

BabelEnconding Similarity consists on a weighted average between word embedding and BERT similarities scaled in multiple languages. Equation 1 shows how BabelEnconding calculates the similarity between words w_1 and w_2 within sentence S_m. In this equation, α and β are the weights given to BERT and Word Embedding similarities, respectively.

$$SIM_{(w_1,w_2)}^{S_m} = \frac{1}{L}\sum_{i=1}^{L} \alpha SIM_{BERT}(w_1^i, w_2^i, S_m) + \beta SIM_{WE}(w_1^i, w_2^i) \quad (1)$$

Our hypothesis is that having similarity information from more languages helps decide on how similar they are. The underlying assumption is that if two words are translated to the same word in other language, they are more likely to be more similar. Translation also helps identifying dissimilarity between words as it can help to disambiguate terms.

Preliminary tests showed that, once both words occur together in the same context, the similarity between words tended to be undesirably high when using just BERT representations. This effect can be attributed to BERT's attention mechanism. Thus, a combination of BERT and fixed word embeddings was designed to alleviate this issue.

4 Experimental Setup

Dataset. The dataset used in our experiments was CoSimLex (Armendariz et al., 2020b) which consists of 340 sentence pairs in English (EN), 112 in Croatian (HR), 111 in Slovene (SL), and 24 in Finnish (FI). Please refer to that paper for details on the annotation methodology.

Languages. The source sentences were translated into the following languages: English (EN), Spanish (ES), Italian (IT), Bosnian (BS), German (DE), Greek (EL), Polish (PL), Portuguese (PT), Russian (RU), Serbian (SR), and Turkish (TR). This choice was made based on the main languages used in Word Sense Disambiguation tasks (Camacho-Collados et al., 2016; Duong et al., 2017; Resnik, 2004; Raganato et al., 2017; Fernández, 2017).

Tools and Resources. The official experiments used Google Translator API[1]. Here, we also report a comparison with Bing Microsoft Translator[2]. The multilingual uncased version of BERT[3] trained on Wikipedias in 102 languages was used. For word embeddings, we used FastText[4] which provides pre-trained embeddings for 157 languages. These embeddings were also trained on Wikipedia.

Evaluation Metrics. The evaluation metrics used to assess the quality of the participating systems measure the correlation between the scores assigned by human annotators and the scores automatically

[1] https://cloud.google.com/translate/

[2] https://www.bing.com/translator

[3] https://github.com/google-research/bert/blob/master/multilingual.md

[4] https://fasttext.cc/docs/en/pretrained-vectors.html

Configuration			Correlation scores Subtask 1				Correlation scores Subtask 2			
Extra Langs.	α	β	EN	HR	FI	SL	EN	HR	FI	SL
None	0.7	0.3	**0.730**$^{(7)}$	0.634	0.607	**0.646**$^{(3)}$	0.615	0.583	0.376	0.559
PT, EL, TR, RU	0.8	0.2	0.683	0.703	0.707	0.617	0.620	0.635	**0.611**$^{(2)}$	0.578
ES, IT, PT, DE	0.6	0.4	0.709	0.735	**0.726**$^{(3)}$	0.525	0.626	0.647	0.571	**0.579**$^{(1)}$
11 Languages	0.7	0.3	0.695	0.716	0.718	0.575	**0.634**$^{(10)}$	**0.658**$^{(1)}$	0.581	0.566
11 Languages	1.0	0.0	0.711	**0.740**$^{(1)}$	**0.726**$^{(3)}$	0.624	0.614	0.632	0.557	0.578
Best Score in SemEval Task3			0.774	0.740	0.772	0.654	0.723	0.658	0.645	0.579

Table 1: Official results for BabelEnconding. Numbers in brackets indicate our ranking.

generated by the participating systems (higher scores are better). In Subtask 1, the uncentered Pearson correlation between was used and, in Subtask 2, the harmonic mean between Pearson and Spearman correlations was used.

Experimental Runs. In order to have a broader evaluation of BabelEnconding, we tested different system configurations and parameters. With the additional runs the goal was answering three questions – (i) *How much each component of BabelEnconding contributes to the overall result?*; (ii) *Do results improve as more languages are added?*; and (iii) *Does the translation mechanism impact the results?*.

5 Results

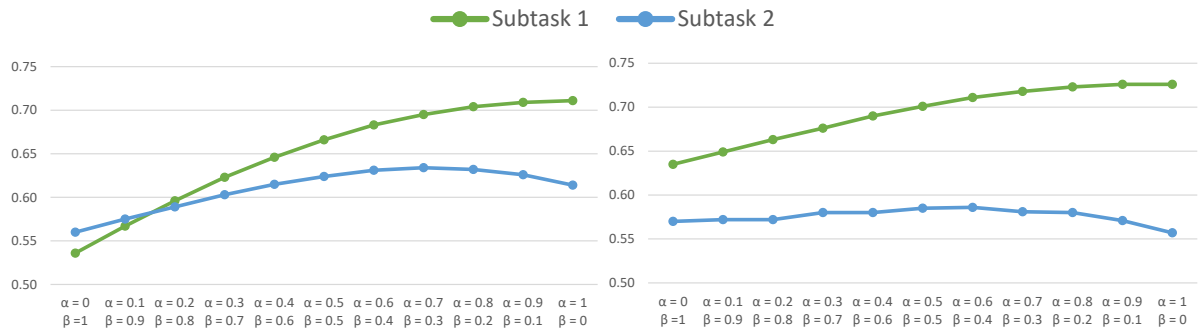
Results for the Official Runs. The system configurations that achieved the best results in the official runs are shown in Table 1. We varied the number of extra languages and the values for α and β. For Subtask 1, English and Slovenian performed better when no additional languages were used in the similarity computation. On the other hand, Croatian and Finnish performed better when all 11 additional languages were used. Moreover, these two languages were benefited when word embeddings were completely removed from BabelEnconding calculation. In Subtask 2, the use of all extra languages or a subset of the 11 languages showed the best results. A combination of BERT and word embeddings also proved to be beneficial for that task. In comparison with other participants, we achieved best results for Croatian, in both subtasks, and for Slovenian in Subtask 2.

Table 2 summarizes the official results for both Subtask 1 and Subtask 2. The column *Average* shows the average of the results achieved by the teams among all languages and the column *Rank* shows the team's position in the ranking. As we can see, our method performed well in both subtasks, being ranked in first place considering the average of all languages.

	Subtask 1						Subtask 2					
TEAM	EN	HR	FI	SL	Average	Rank	EN	HR	FI	SL	Average	Rank
BabelEncoding	0.730	**0.740**	0.726	**0.646**	**0.710**	1	0.634	**0.658**	0.611	**0.579**	**0.620**	1
Team 1	**0.774**	0.634	0.745	0.605	0.689	2	0.437	0.397	0.357	0.345	0.384	10
Team 2	0.768	0.594	**0.772**	0.583	0.679	3	0.695	0.385	0.341	0.485	0.476	6
Team 3	0.754	0.664	0.626	0.648	0.673	4	0.715	0.545	**0.645**	0.573	0.619	2
Team 4	0.712	0.681	0.574	**0.654**	0.655	5	0.695	0.616	0.255	0.510	0.519	5
Team 5	0.754	0.616	0.360	0.560	0.572	6	0.720	0.565	0.354	0.483	0.530	4
Team 6	0.738	0.440	0.546	0.512	0.559	7	-	-	-	-	-	-
Team 7	0.529	0.531	0.399	0.510	0.492	8	-	-	-	-	-	-
Team 8	0.042	0.587	0.671	0.603	0.475	9	0.647	0.402	0.289	0.516	0.463	7
Team 9	0.721	0.416	0.025	0.624	0.446	10	-	-	-	-	-	-
Team 10	0.544	0.374	0.389	0.328	0.408	11	**0.723**	0.613	0.597	0.487	0.605	3
Team 11	-	-	-	-	-	-	0.573	0.402	0.289	0.516	0.445	8
Team 12	-	-	-	-	-	-	0.340	0.338	0.454	0.411	0.385	9

Table 2: Official Results for Subtask 1 and Subtask 2 for all participating teams. Bold indicates the best result for the given language. (Armendariz et al., 2020a)

How much each component of BabelEnconding contributes to the overall result? In order to assess the contribution of the components of BabelEnconding, we performed experiments varying the parameters α (which scales the contribution of BERT similarity), and β (which weighs the importance of word embedding similarity). As a general tendency, increasing α values tends to produce better correlation results, especially in Subtask 1. However, when the word embeddings component is removed (*i.e.,* β=0), results tend to get worse, mainly in Subtask 2. Figure 2 shows the results for English and Finnish. The curves in (a) represent the typical case, which was found in English, Croatian, and Slovenian. The results for Finnish (b) in Subtask 2 followed a different pattern, in which evaluation scores are not affected by the presence of BERT on similarity computation. We believe this happened because Finnish is an agglutinative language, and since BERT's tokenization process uses Byte Pair Encoding, it tends to split Finnish words in too many tokens (Virtanen et al., 2019) yielding to poorer word representations.

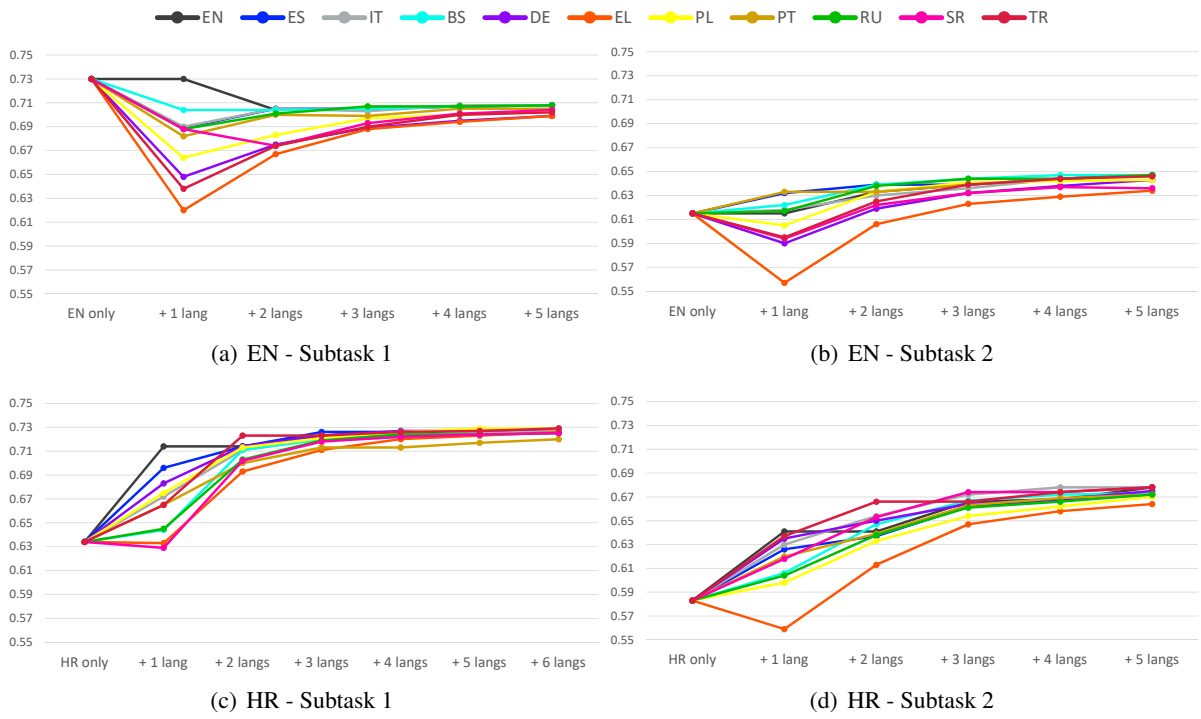


(a) English (b) Finnish

Figure 2: Correlation scores with human judges for different values of α and β in BabelEnconding

Do results improve as more languages are added? In order to evaluate the benefits of multilingualism, we performed an experiment in which the performance using only the source language (*i.e.,* the language of the original sentence) is compared to the performance when more languages are incrementally added. Figure 3 shows the results for this experiment for the datasets in English and Croatian. The first set of points on the plot mark the case in which only the original language was used. The second set, shows the scores when each of 11 possible languages were added. From the third set of points onward, we kept the language(s) that brought the biggest gain and added one more. We repeated this process until the addition of a new language ceased to bring improvements. The combination of multiple languages was beneficial for Croatian, in both subtasks, and for English in Subtask 2. In Croatian, the addition of one language improved results in 9 out of 11 possible languages. The exceptions were Greek and Serbian, in which cases, the scores remained the same. By adding English, the score increased by eight percentage points. By adding further languages, the improvement was smaller but steady until it reached a plateau with six additional languages.

Does the translation mechanism impact the results? In order to evaluate the impact of different translation engines on BabelEnconding, we compared the performance of Google Translator and Bing Microsoft Translator. The four original datasets were translated into the 11 languages using both engines. Then, the translated datasets were used to perform the contextual similarity tasks with the same algorithm configuration (all languages considered, $\alpha = 0.7$ and $\beta = 0.3$). The results are shown in Figure 4. Google Translator outperforms Bing in both subtasks for all languages. The superior performance of Google Translator is in line with the findings from other recent works – Marzouk and Hansen-Schirra (2019) evaluated translations from of German to English and found that Google presented the best results and Way et al. (2020) evaluated the translation of technical texts and found that, in most of cases, the translations provided by Google were better. Intuitively, better translations yield better contextual similarity and that was confirmed here.



(a) EN - Subtask 1

(b) EN - Subtask 2

(c) HR - Subtask 1

(d) HR - Subtask 2

Figure 3: The effects of adding more languages to the similarity computation in BabelEnconding. The numbers reflect the correlation with the human-generated similarity scores.

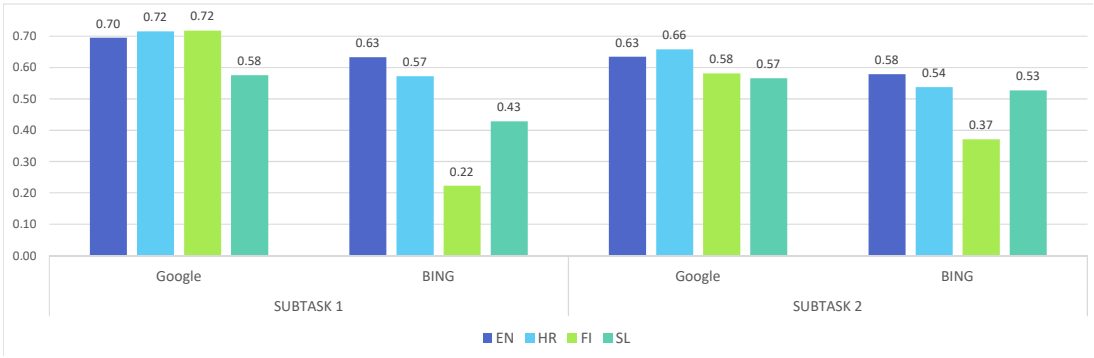


Figure 4: Correlation scores with human judges for different translation engines in BabelEnconding

6 Conclusion

In this paper, we described our system submitted to SemEval-2020 Task 3. We designed an approach that relies on translation and multilingual language models in order to compute the contextual similarity between pairs of words. The key idea is that having similarity information from different languages may help decide on how similar the words are. Our system achieved competitive results in both subtasks, being ranked among the top-3 in most runs.

In these preliminary experiments, we could not establish in which cases more languages are helpful and we leave it as future work. Additionally, we are interested in understanding which factors contribute to improvement in the results – whether it is the amount of data used for training the language models or individual features of the language.

Acknowledgement. This work was partially supported by CNPq/Brazil and by CAPES Finance Code 001.

References

Carlos S. Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Marko Robnik-Šikonja, Ivan Vulić, and Mohammad Taher Pilehvar. 2020a. SemEval-2020 task 3: Graded word similarity in context (GWSC). In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Carlos S. Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. 2020b. CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5878–5886, Marseille, France, May. European Language Resources Association.

Carmen Banea and Rada Mihalcea. 2011. Word sense disambiguation with multilingual features. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Marine Carpuat. 2013. NRC: A machine translation approach to cross-lingual word sense disambiguation (SemEval-2013 task 10). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 188–192, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Bharath Dandala, Rada Mihalcea, and Razvan Bunescu. 2013. Multilingual word sense disambiguation using wikipedia. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 498–506.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, June.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2017. Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 894–904.

Alejandro Moreo Fernández, Andrea Esuli, and Fabrizio Sebastiani. 2016. Distributional correspondence indexing for cross-lingual and cross-domain sentiment classification. *Journal of artificial intelligence research*, 55:131–163.

Andrés Duque Fernández. 2017. *Word sense disambiguation in multilingual contexts*. Ph.D. thesis, UNED. Universidad Nacional de Educación a Distancia (España).

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Yoav Levine, Barak Lenz, Or Dagan, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. Sensebert: Driving some sense into bert. *arXiv preprint arXiv:1908.05646*.

Timo Lüddecke, Alejandro Agostini, Michael Fauth, Minija Tamosiunaite, and Florentin Wörgötter. 2019. Distributional semantics of objects in visual scenes in comparison to text. *Artificial Intelligence*, 274:44–65.

Shaimaa Marzouk and Silvia Hansen-Schirra. 2019. Evaluation of the impact of controlled language on neural machine translation compared to other mt architectures. *Machine Translation*, 33(1-2):179–203.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Matthew E Peters, Mark Neumann, IV Logan, L Robert, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2018. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121*.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167.

Philip Resnik. 2004. Exploiting hidden meanings: Using bilingual text for monolingual annotation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 283–299. Springer.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proc. of AAAI*.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Ivan Vulic, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-simlex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. *CoRR*.

Shirui Wang, Wenan Zhou, and Chao Jiang. 2020. A survey of word embeddings based on deep learning. *Computing*, 102(3):717–740.

Andy Way, Rejwanul Haque, Guodong Xie, Federico Gaspari, Maja Popović, and Alberto Poncelas. 2020. Rapid development of competitive translation engines for access to multilingual covid-19 information. In *Informatics*, volume 7, page 19. Multidisciplinary Digital Publishing Institute.