Approaching Semantic Text Similarity with Hybrid Methods: a Case Study on French

Antonella Fadda[1], Rémi Cardon[2], Natalia Grabar[3], Thomas François[4]

[1]FIAL, UCLouvain – antonella.fadda@student.uclouvain.be

[2]Cental, UCLouvain – remi.cardon@uclouvain.be

[3]UMR STL, UdLille – natalia.grabar@univ-lille.fr

[4]Cental, UCLouvain – thomas.francois@uclouvain.be

Abstract (in English)

The difficulty in understanding texts is a daily struggle for many people. To overcome this problem, Natural Language Processing (NLP) offers various solutions, namely text simplification. The main difficulty in developing systems for text simplification is the lack of resources, such as parallel corpora or lexicons. One common approach for parallel corpora development is extraction of sentences that share the same meaning, from comparable corpora. Doing so requires evaluating the semantic similarity between sentence pairs. In this article, we propose to investigate this task in the light of the recent developments in NLP. Concretely, we will work on the French language, using two corpora : DEFT'20 and CLEAR. DEFT'20 is a French corpus containing 1,010 sentence pairs annotated with their degree of similarity on a 0-5 scale. CLEAR is a French comparable biomedical corpus made for text simplification out of three different sources, Wikipedia/Vikidia, drug leaflets, and medical literature summaries. We report on experiments with state of the art language models for French (general such as CamemBERT and FlauBERT) and with classic feature-based machine learning approaches (e.g. Random Forest with similarity measures such as Manhattan distance, Levenshtein distance, Dice coefficient, etc.). As we observe that the top-performing systems of the DEFT 2020 campaign on the task achieve similar results as the language models in isolation, we closely analyze the strengths and weaknesses of the two approaches in order to identify how complementary they are. We evaluate our experiments in two ways: (1) by their performance on the DEFT'20 corpus and (2) by their ability to identify parallel sentences from the CLEAR comparable corpus.

Keywords: semantic text similarity ; sentence alignment ; natural language processing

1. Introduction

The difficulty in understanding texts is a daily struggle for many people in different situations, such as: reading official documents, communicating with specialists from a given area (medicine, banking, law…), searching for information online. This problem is also being addressed by the Natural Language Processing (NLP) methods. Different aspects of this research question can be distinguished: definition of the document or sentence difficulty, acquisition of lexicon for the simplification, automatic text simplification. One way to address Automatic text simplification is to build and use sets with parallel and aligned sentences, where the meaning of the sentence is quite close but the language register is different: complex or technical sentences are aligned with the corresponding simple or simplified sentences, like in the example (1) below.

(1) Sentence 1: *Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation.*
Sentence 2: *Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou à ralentir le transit intestinal.*

In this work, we propose to address the task of building such resources for French. This is the aim of a task called Semantic Textual Similarity (STS). The task consists in evaluating the level of semantic relatedness of two given sequences of text. We explore the literature on the subject in Section 2. In Section 3, we describe the data we use for our experiments, which are STS corpora. Section 4 describes the methodology we use for our various experiments. In Section 5, we report the results of the experiments and propose an error analysis of our best approaches. Finally, we conclude in section 6.

2. Related Work

Semantic Textual Similarity (STS) is a line of research in NLP that has largely been explored, as many NLP applications can benefit from it. STS addresses the following question: for any two given textual sequences (documents, paragraphs, sentences…), to what extent do they express the same meaning? This is mostly answered on a continuous Lickert scale from 0 (two unrelated sequences) to 5 (the two sequences express the same meaning). The task has extensively been explored in the context of three SemEval shared tasks, in 2015 (Xu et al., 2015), 2016 (Bethard et al., 2016) and 2017 (Cer et al., 2017). A recent survey (Chandrasekaran and Mago, 2021) establishes four types of methods:

1. **Knowledge-based methods**. Those methods derive similarity information from existing structured databases, such as WordNet (Miller, 1995) for example;
2. **Corpus-based methods**. They consist of computing similarity on a meaning representations based on the distributional hypothesis. Examples of building such representations are Latent Semantic Analysis (Landauer and Dumais, 1997) or more recent embedding methods, such as Word2Vec (Mikolov et al., 2013) or BERT (Devlin et al., 2018) models. Those techniques require large corpora or pre-trained representations;
3. **Deep neural networks-based methods.** This approach amounts to training neural networks for the task. Various variations on well-known architectures like CNNs (LeCun et al., 2015) or LSTMs (Hochreiter and Schmidhuber, 1997) have been proposed for the STS task. Those techniques require a large amount of labeled data for training the models;
4. **Hybrid methods**. Those methods typically combine corpus-based or deep neural networks-based methods with knowledge-based methods (Camacho-Collados et al., 2015; Ruas et al., 2015).

According to this survey, hybrid models usually compensate for the shortcomings of one method by incorporating other methods. Hence the performance of hybrid methods is comparatively high. In our work, we want to investigate a hybrid approach on a French-language STS dataset.

3. Data

For our experiments we use an existing French corpus made for semantic similarity (Cardon and Grabar, 2020). This corpus was used for an STS shared task within the DEFT 2020 challenge (Cardon et al., 2020). It comprises a total of 1,010 sentence pairs, and is split into a training set (600 sentence pairs) and a test set (410 sentence pairs). The corpus is composed of sentences extracted from "Wikipedia and Vikidia pages on various subjects [...] as well as health-related content such as drug inserts [...] and Cochrane summaries." (Cardon et al., 2020, p. 2). Each sentence pair was annotated by five annotators, who assigned them a score on a Lickert scale from 0 to 5. A value of 0 indicates completely dissimilar sentences, while a value of 5 represents sentences with identical meaning. Besides the individual score given by each annotator, the average of the scores is also present. We use the latter as the target value.

From the CLEAR corpus (Grabar and Cardon, 2018), we extracted 500 pairs, 100 of which are annotated as equivalent (positive) and the other 400 as not equivalent (negative). The binary annotation makes the task a binary classification task on this corpus. This corpus enables us to do two things: (1) explore the STS task in both regression and classification settings and (2) check the validity of our approach to mine parallel sentences from a comparable corpus.

4. Methodology

In this section, we introduce the methodology for two series of experiments that we run. One leverages the DEFT'20 dataset (section 4.1) and the other one the CLEAR corpus (section 4.2).

4.1 Regression on the DEFT'20 dataset

We proceed with three sets of experiments with regression. For each, we use the train and test splits of the DEFT corpus as they are provided by the organizers. One set of experiments consists in comparing the performance of two language models for embedding French sentences: CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020). Note that in what follows, when we refer to both at the same time, we use the name *BERT. To do so, we train a Random Forest regression model on top of the concatenation of the embeddings of each sentence for a given pair.

As a first step we tokenize the sentence with the *BERT tokenizer. Special tokens such as [CLS] (classification) and [SEP] (separator) are added. Due to architectural constraints proper to *BERT, the max length has been set to 512 tokens. Padding is applied to ensure uniformity. We generate *BERT embeddings for each tokenized sentence, and the hidden states of the last layer are extracted. We calculate the average of the hidden states for all words in the sentence, considering the attention mask associated with each sentence in order to avoid including padding tokens in the calculation. Only the best-performing of the two language models used in this series of experiments is kept for the other experiments.

Another set of experiments is the study of various classic similarity measures. The similarity measures we choose are those that were also used in the DEFT 2020 campaign and that

produced good results. On the one hand, we include similarity measures calculated on the embeddings, such as Manhattan distance, Euclidean distance, cosine similarity and dissimilarity, Bray-Curtis dissimilarity (Bray and Curtis, 1957), and Ochiai coefficient (Ochiai, 1957). On the other hand, other measures were calculated comparing the two strings of characters, and they include the Sorensen Dice coefficient (Dice, 1945),Jaccard distance (Jaccard, 1912), the Levenshtein distance (Levenshtein, 1966), the Qgram coefficient (Ukkonen, 1992), the number of words in common, the length of each sentence and the absolute difference between two sentences. For these measures, calculated directly from the sentence and not from its latent representation of the embeddings, we performed a pre-processing of the sentences, which involves the removal of stopwords using the NLTK package and the removal of all non-alphanumeric characters. Like for the first set of experiments, we represent the sentence pairs.

In order to find the best combination of features, in this set of experiments we test all the possible subsets of similarity measures. Since there are 14 measures, a total of 16,383 Random Forest models are trained. Referring to the Spearman's correlation measure calculated for every Random Forest model, we found the best combination of measures.

The third set of experiments is to use the best set of measures we found in combination with the best of the two language models. Indeed, due to a lack of computing power, we are not able to re-train more than sixteen thousand Random Forests that include the embeddings and all possible combinations of measures. We therefore chose to limit this last step to the best combination of measures. We concatenate the embeddings to the feature vectors produced by the best combination of similarity measures and train a Random Forest model on this.

As explained in the Data section, the label of each sentence pair is the average of the scores given by the annotators. This label is a value within a range from 0 to 5. This score being on a continuous numerical scale, we use a regressor to model it. Moreover, since the Random Forest was the model that had the best results in the DEFT 2020 campaign, we decided to use it in the current experiment as well.

We use the RandomForestRegressor class from the Scikit-learn library (Pedregosa et al., 2011), version 1.3.2. As for the hyperparameters, we set the number of estimators to 500 and a random state of 0 for reproducibility, and kept the other hyper-parameters at their default settings. This decision is supported by multiple trials performed that attest to its effectiveness. Variations in the 'max depth' and 'max features' parameters had no significant impact on the results. The 'max depth' parameters were tested with values of None (default), 50, 100 and 150, while for 'max features' the choices included sqrt and log2.

After calculating the embeddings and the similarity measures, we proceed to train the Random Forest regressor using the hyperparameters specified above. We evaluate our experiments using the Spearman correlation.

4.2 Classification on sentences from the CLEAR corpus

As a last series of experiments, we apply our best models for each set of experiments, trained on the DEFT20 corpus, to the labeled sentence pairs from the CLEAR corpus. Since our models are trained on a regression task, whereas those pairs of sentences are tailored for a binary classification task (either they are similar or they are not), we test different thresholds. If the score is higher than the threshold, the pair is marked as similar, if the score is equal to

or lower than the threshold, the pair is considered as not similar. We run this set of experiments with all threshold values between 0.1 and 4.9 in steps of 0.1. We evaluate those experiments with accuracy.

5. Results and Discussion

We present the results along three lines: processing of the data with Regression models (Section 5.1), processing of the data through Classification (Section 5.2), and an Error analysis (Section 5.3).

5.1 Regression

In this section, we report the results obtained with the Regression models. Our evaluation makes use of Spearman's correlation and is structured in three parts: (1) the results obtained from the Random Forest trained on embeddings only (Table 1), (2) the results obtained from the similarity measures and (3) finally the results obtained from the concatenation of the best combination of measures and embeddings. As far as the results of the combinations of measures are concerned, we decided to present only the combination that obtained the very best results. This choice is due to the fact that we do not observe a significant divergence between the Spearman correlations of the various combinations.

Table 1. Results of the RF training on embeddings only

DATA	Spearman
Camembert embeddings	0.76
Flaubert embeddings	0.74

However, specific combinations can be identified that demonstrate suboptimal performance relative to others. These include the length of the two sentences and the Ochiai coefficient. In particular, these metrics consistently rank in the bottom 10% of the results, each with a frequency of 0.08. They also consistently remain among the least effective metrics in the bottom 20%, where the length of the two sentences has a frequency of 0.16 and the Ochiai coefficient has a frequency of 0.17.

The frequencies of the most effective measures follow a similar pattern. Specifically, the Manhattan distance, the absolute difference between the length of the two sentences, the Levenshtein distance, and the Qgram similarity all rank in the top 10%. The Manhattan distance is the most present with a frequency of 0.83, followed by the absolute difference with 0.74, the Levenshtein distance with a frequency of 0.69 and the Qgram similarity with 0.69. Within the 20% range, the ranking does not change: the Manhattan distance is still the most frequent with 0.72 followed by the absolute difference with 0.70. The Levenshtein distance has a frequency of 0.66 and the Qgram similarity follows with 0.63.

Our experiments indicate that the best combination is composed of the following measures: Manhattan distance, Euclidean distance, Sorensen-Dice coefficient, Common words, length of first sentence, length of second sentence, absolute difference between the two sentences, Levenshtein distance, Qgram similarity with a Spearman's correlation of 0.85. If we round to the second digit, only after the 156th combination does the correlation drop to 0.84.

The results of concatenation of embeddings and the best combination of measures has a Spearman's correlation of 0.88, which is 0.11 above the best performance obtained by the DEFT 2020 participants. It is also 0.04 above the model trained only on similarity measures, and 0.12 above the model trained on CamemBERT embeddings.

5.2 Classification

We report results on three Classification models, based on the results that are described in the previous section. The three models are: (1) the model trained on CamemBERT embeddings only, (2) the best performing model trained on similarity measures only, and (3) the model that is trained on the concatenation of CamemBERT embeddings and the best performing combination of similarity measures. We apply those models to the 500 sentence pairs (100 positive, 400 negative) from CLEAR and report the accuracy for each threshold. We report only results thresholds between 1.5 and 3.5 for readability, as we would not expect a good performing threshold to fall outside of this scope. We also report the results for the minimum (0.1) and maximum (4.9) thresholds.

The results are displayed in Table 2. Interestingly, we can see that the best accuracy is obtained by the model trained on measures only, with a score of 97.2 at threshold 2.2. The CamemBERT+measures model is not far behind as it peaks at 96, at thresholds 2.9 and 3. The model including only CamemBERT embeddings performs the worst. These results suggest that it cannot be leveraged to produce a threshold that distinguishes the two classes: its highest score (accuracy of 80%) is at the maximum threshold (4.9), which means that everything is predicted as non similar (0). As 80% of examples are negative examples, this model cannot do better than a system that would assign everything to the negative class, despite its Spearman's correlation that suggests it would achieve good results.

Table 2. Accuracy results for the three retained methods on the binary classification task

Reference similarity threshold	**CamemBERT**	**Measures**	**CamemBERT+Measures**
0.1	20.0	27.4	20.0
1.5	20.4	87.8	53.8
1.6	21.2	90.8	58.6
1.7	21.6	94.6	62.2
1.8	23.0	96.6	68.2
1.9	24.0	96.4	70.0
2.0	25.6	97.0	73.8
2.1	26.8	97.0	78.2

2.2	29.6	**97.2**	82.4
2.3	32.2	96.8	87.0
2.4	34.8	96.6	90.6
2.5	36.4	96.4	93.6
2.6	40.2	95.8	94.8
2.7	44.4	95.6	95.4
2.8	47.8	95.2	95.8
2.9	51.6	94.4	**96.0**
3.0	55.0	92.4	**96.0**
3.1	62	91.4	95.0
3.2	68.2	90.4	93.2
3.3	73.2	89.2	91.4
3.4	75.6	88.4	89.6
3.5	78.8	87.0	88.6
4.9	**80.0**	80.0	80.0

5.2 Error Analysis

We manually looked at the errors produced by two classification models, using the best threshold for each, namely 2.2 for the similarity measure-based one and 2.9 for the one combining CamemBERT and the measures.

Those manual observations are quite convenient to make due to the low number of errors : the first model produced 11 false negatives (equivalent pairs that are considered unrelated) – which places its recall at 0.89 for the positive class – and 3 false positives (unrelated pairs that are considered equivalent) – which places its precision at 0.97 for the positive class. The second model produced 13 false negatives – recall at 0.87 for the positive class – and 7 false positives – precision at 0.93 for the positive class.

Ten identical false negatives were produced by both models. There seems to be a common pattern, as the sentences in those pairs differ by the way they organize the delivery of the message. For instance, the voice of the verb (passive vs. active) or an impersonal form vs. a sentence that addresses the reader. Below we show one example of each case:

(2) Sentence 1: *Éviter la prise de boissons alcoolisées et de médicaments contenant de l'alcool.*

Sentence 2: *La prise d'alcool est formellement déconseillée pendant la durée du traitement.*

(3) Sentence 1: *Il est préférable d'utiliser d'autres traitements ayant un profil de sécurité bien établi pendant l'allaitement, particulièrement chez le nouveau-né ou le prématuré.*
Sentence 2: *Moex est déconseillé aux femmes qui allaitent et votre médecin pourrait choisir un autre traitement si vous souhaitez allaiter, surtout si votre enfant est un nouveau-né ou un prématuré.*

Interestingly, the few remaining false negatives that have been produced by only one of the two models follow the same pattern (one by the first model, three by the other one), indicating that some of those cases are recognized by the models. This tends to suggest that adding some information on the syntactic organization of the constituents (maybe with their semantic roles) would be the next step to improve the classification.

Regarding the false negatives, they are typically long sentences that share words in common even though they do not convey the same message. One example is shown below (we put in bold the phrases present in the two sentences that we suspect to be the origin of the error):

(4) Sentence 1: *Tous les essais contrôlés randomisés (ECR) ou quasi-ECR de l'adjonction* ***de corticostéroïdes*** *dans le traitement des* ***nouveau-nés atteints de méningite bactérienne****.*
Sentence 2: *Est-ce que l' utilisation* ***de corticostéroïdes*** *adjuvants chez les* ***nouveau-nés atteints de méningite bactérienne*** *réduit le risque de décès et la possibilité d'avoir des séquelles neurodéveloppementales ?*

This also leads to the next step being adding information about syntactic organization and semantic roles: while it would enable recognizing the similarity of seemingly different phrases in false negatives, it would enable recognizing the difference of seemingly identical phrases in false positives.

6. Conclusion

The results presented in this paper highlight the effectiveness of concatenating embeddings and similarity measures for sentence similarity assessment. In particular, the results obtained show significant improvements over the results of the DEFT 2020 campaign. Converting regression models into classification models using a threshold yielded interesting results. Two approaches (the measures-only approach and the hybrid approach) perform strongly in the classification mode, with thresholds that are different (around 2 for the first approach and around 3 for the second approach).

As a limitation, it is important to state that, due to our inability to compute the 16,383 combinations of measures with the use of embeddings, we cannot know whether we found the optimal combination for our CamemBERT+measures model. Indeed, some sets of measures combined with embeddings might prove more effective than the highest performing set of measures on its own. Nonetheless, the end results we obtained show that our methodology is

efficient for the semantic textual similarity task, while not requiring large amounts of resources dedicated for the task or computational power.

In future work, we will plan to enrich the model with additional representations coming from syntactic and semantic knowledge. Typically, the active and passive voice in sentences may be better taken into account. We also intend on applying our method on existing comparable corpora, in order to produce better resources for tasks that benefit from monolingual parallel corpora, such as works on paraphrase, or tasks like automatic text simplification.

References

Bethard S., Carpuat, M., Cer, D., Jurgens, D., Nakov, P., and Zesch, T. (2016). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, edition.

Bray J. R., Curtis J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4), 325–349

Camacho-Collados, J., Pilehvar, M. T., and Navigli. R. (2015). NASARI: A novel approach to a semantically aware representation of items. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 567–577

Cardon, R. and Grabar, N. (2020). A French Corpus for Semantic Similarity. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6889–6894, Marseille, France. European Language Resources Association.

Cardon, R., Grabar, N., Grouin, C., and Hamon, T. (2020). Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition).* Atelier DÉfi Fouille de Textes, pages 1–13, Nancy, France. ATALA et AFCP.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Chandrasekaran, D. and Mago, V. (2021). Evolution of Semantic Similarity—A Survey. *ACM Computational Surveys* 54, 2, Article 41 (March 2022), 37 pages.

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

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297-302.

Grabar, N. and Cardon, R. (2018). CLEAR – Simple Corpus for Medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Hochreiter, S., and Schmidhuber, Jürgen. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Jaccard P. (1912). The Distribution of the Flora in the Alpine Zone.1. *New Phytologist* 11 (2): 37–50.

Landauer, T. K. and Dumais S. T.. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 2 (1997), 211.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux; B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature*, 521 (7553), pp.436-444.

Levenshtein V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10 (February): 707.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, E., Seddah, D., and Sagot, B. (2020). CamemBERT : A Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G. and Dean. J. (2013). Efficient estimation of word representations in vector space. *Arxiv Preprint Arxiv:1301.3781* (2013).

Miller, G.A. (1995). WordNet: A lexical database for English. *Communications of the ACM* 38, 11 (1995), 39–41.

Ochiai A. (1957). Zoogeographical Studies on the Soleoid Fishes Found in Japan and Its Neighhouring Regions-II. *NIPPON SUISAN GAKKAISHI* 22 (9): 526–30.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825--2830.

Terry Ruas, William Grosky, and Akiko Aizawa. 2019. Multi-sense embeddings through a word sense disambiguation process. *Expert Systems with Applications* 136 (2019), 288–303

Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R., and Rezende S. O. (2019). Knowledge-enhanced document embeddings for text classification. *Knowledge-based Systems* 163 (2019), 955–971.

Ukkonen E. (1992). Approximate String-Matching with q-Grams and Maximal Matches. *Theoretical Computer Science* 92 (1): 191–211.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit,, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin. I. (2017). Attention is all you need. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*.

Xu, W., Callison-Burch, C., and Dolan, B. (2015). SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.