ExB Themis: Extensive Feature Extraction from Word Alignments for Semantic Textual Similarity

Christian Hänig, Robert Remus, Xose De La Puente
ExB Research & Development GmbH
Seeburgstr. 100
04103 Leipzig, Germany
{haenig, remus, puente}@exb.de

Abstract

We present *ExB Themis* – a word alignment-based semantic textual similarity system developed for SemEval-2015 Task 2: Semantic Textual Similarity. It combines both string and semantic similarity measures as well as alignment features using Support Vector Regression. It occupies the first three places on Spanish data and additionally places second on English data. *ExB Themis* proved to be the best multilingual system among all participants.

1 Introduction

Semantic Textual Similarity (STS) is the task of measuring the degree of semantic equivalence of a sentence pair and is applicable to problems in Machine Translation and Summarization among others (Agirre et al., 2012). STS has drawn a lot of attention in the last few years leading to the availability of multilingual training and test data and to the development of a variety of approaches. These approaches fall broadly into three categories (Han et al., 2013):

Vector space approaches: Texts are represented as bag-of-words vectors and a vector similarity – e. g. cosine – is used to compute a similarity score between two texts (Meadow et al., 1992).

Alignment approaches: Words and phrases in two texts are aligned and the quality or coverage of the resulting alignments are used as similarity measure (Mihalcea et al., 2006; Sultan et al., 2014).

Machine Learning approaches: Multiple similarity measures and features are combined using supervised Machine Learning (ML). This approach relies on the availability of training data (Bär et al., 2012; Šarić et al., 2012).

ExB Themis combines advantages of all three categories: we implemented a complex alignment algorithm focusing on named entities, temporal expressions, measurement expressions and dedicated negation handling. Unlike other alignment-based approaches, we extract a variety of features to better model the properties of alignments instead of providing only one alignment feature (see Section 4.1).

Moreover, we employ a variety of similarity measures based on strings and lexical items (see Section 4.2). Our system integrates two well-known language resources – WordNet[1] and ConceptNet (Speer and Havasi, 2012). Additionally, it uses word embeddings to cope with data sparseness and the insufficiency of overlaps between sentences.

Finally, we train a Support Vector Regression (SVR) model using these features (see Section 5).

2 Preprocessing

Our text preprocessing comprises tokenization, case correction (e. g. *US Flying Surveillance Missions to Help Find Kidnapped Nigerian Girls* is corrected to *US flying surveillance missions to help find kidnapped Nigerian girls*), unsupervised part-of-speech (POS) tagging based on SVD2 (Lamar et al., 2010),

[1]English: we use the one described by Miller (1995); Spanish: we use the one presented in (González-Agirre et al., 2012).

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 264–268,
Denver, Colorado, June 4-5, 2015.

supervised POS tagging using the Stanford Maximum Entropy tagger[2] as well as lemmatization using Stanford CoreNLP[3] for English and IXA Pipes[4] for Spanish. We also identify measurements (e. g. *55.8 g/mol*) and temporal expressions (e. g. *last week*), data set-specific stop words (e. g. *A close-up of* for *images* dataset) using in-house algorithms as well as named entities as described by Hänig et al. (2014) and their titles (e. g. *President Barack Obama*).

3 ExB Themis Alignment

Our word alignment is direction-dependent and not restricted to one-to-one alignments. Different mapping types are distinguished and handled differently during feature extraction (see Section 4.1). We use the same type labels as provided by the organizers for the third subtask (interpretable STS) of this task (Agirre et al., 2015): *EQUI* denotes semantically equivalent chunks, oppositional meaning is labeled with *OPPO*, *SPE1/2* denote similar meaning of the chunks, but the chunk in sentence 1/2 is more specific than the other one. *SIM* and *REL* denote similar and related meanings, respectively. *ALIC* is not used, because our algorithm is not restricted to one-to-one alignments. Finally, all unaligned chunks are labeled with *NOALI*.

Similar to Sultan et al. (2014), our alignment process follows a strict chronological order:

Named entities are aligned to each other. Because we did not observe text pairs with possibly ambiguous name alignments (e. g. *Michael* in one text and both *Michael Jackson* and *Michael Schumacher* in the other) in the training data, we simply aligned all name pairs that share at least one identical token.

Normalized temporal expressions are aligned iff they denote the same point in time or the same time interval (e. g. *14:03* and *2.03 pm*).

Measurement expressions are aligned iff they express the same absolute value (e. g. *$100k* and *100.000$*).

Arbitrary token sequence alignment consists of multiple steps and is very time consuming[5]. We apply a high precision test for identical sequences based on Sultan et al. (2014): Our test uses synonym-lookups and ignores case information, punctuation characters and symbols. This enables us to match expressions like *long term* and *long-term*[6]. If one of both sequences consists of exactly one all-caps-token then we test if it is the acronym of the other sequence (e. g. *US* and *United States*).

We used WordNet and ConceptNet[7] to obtain information about synonymy, antonymy and hypernymy and equip the resulting alignments with the corresponding type. We additionally created a small database containing high-frequency synonyms (e. g. *does* and *do*), antonyms (e. g. *doesn't* and *does*) and negations (e. g. *don't*, *never*, *no*).

Negations can significantly effect the semantic similarity of two sentences (e. g. *You are a Christian.* vs. *Therefore you are not a Christian.*). Therefore, we explicitly model negations in our alignment. Some negations are handled during arbitrary token sequence alignment. We resolve the scope of all remaining negations using co-occurrence analysis: if exactly one of both neighboring tokens $w_{n-1}^{1/2}$ and $w_{n+1}^{1/2}$ is already aligned then the negation $w_n^{1/2}$ is attached to it and we inverse the alignment type (e. g. *EQUI* becomes *OPPO* and vice versa). If both neighboring tokens are aligned then we pick the one contained in the co-occurence out of $\left\langle w_{n-1}^{1/2}, w_n^{1/2} \right\rangle$ and $\left\langle w_n^{1/2}, w_{n+1}^{1/2} \right\rangle$ yielding the highest co-occurrence significance score.

Remaining content words are aligned using cosine similarity on word2vec vectors (Mikolov et al., 2013). Analogously to Han et al. (2013), we align each content word to the content word of the other sentence with the same POS tag that yields the highest similarity score. To prevent weak alignments, we reject alignments with a similarity less than $1/3$.

[2] nlp.stanford.edu/software/tagger.shtml
[3] nlp.stanford.edu/software/corenlp.shtml
[4] ixa2.si.ehu.es/ixa-pipes/

[5] Therefore, we restrict ourselves to a maximum of 5 tokens.
[6] A similar method was described by Han et al. (2013).
[7] From ConceptNet we only imported synonyms.

4 Feature Extraction

Some approaches to STS relying on word alignment are unsupervised and extract a defined score based on the alignment process (e. g. proportions of aligned content words (Sultan et al., 2014)), others extract a single feature from the alignment and use it along with other features to train a regression model (e. g. align-and-penalize approach (Han et al., 2013; Kashyap et al., 2014)).

Unlike these approaches, we extract 40 features from our alignment (see Section 4.1) to (a) build a complex model that is capable of modeling phenomena like alignments of different types and negations, and (b) not be forced to combine alignment properties arbitrarily.

We additionally extract 51 non-alignment features (see Section 4.2) leading to a total of 91 features.

4.1 Alignment Features

To encode the properties of a set of alignments A of sentences s_1 and s_2 as comprehensive as possible, we extract the following features[8]:

Proportion features describe the ratio of aligned words of a specified group with respect to all words of that group (Sultan et al., 2014)[9]: $prop_{group} = \frac{2 \cdot prop^1_{group} \cdot prop^2_{group}}{prop^1_{group} + prop^2_{group}}$ with $prop^{1/2}_{group} = \frac{|\{i:[\exists j:(i,j) \in A_{group}] \text{ and } w_i^{1/2} \in C\}|}{|\{i:w_i^{1/2} \in C\}|}$ where C is the set of all content words. We extract these features for alignments of type *EQUI*, *OPPO*, *SPE1/2*, *REL*[10] and *NOALI* (5 features).

Frequency features are encoded in binary format.

We encode frequencies of alignments of type *OPPO* (3 features), *SPE1/2* (3), *REL* (3) and *NOALI* (5). We also encode the frequency of unaligned negations with 3 features.

UMBC align-and-penalize features: We also include two features[11] based on Han et al. (2013): we use their T as it is and integrated a simplified version of P' with $P_i^A = \frac{\sum_{\langle t,g(t)\rangle \in A_i} (1+w_p(t))}{2 \cdot |s_i|}$ and $P_i^B = \frac{|\langle t,g(t)\rangle \in B_i|}{2 \cdot |s_i|}$ (2 features).

All proportion features, binary frequency features of *REL*-alignments, unaligned content words and unaligned negations were additionally computed and extracted for nouns only (16 features).

4.2 Non-Alignment Features

We use a variety of non-alignment features:

UKP: We use several features described in Bär et al. (2012): longest common substring (1 feature), longest common subsequence (1), longest common subsequence with and without normalization (2), greedy string tiling (1), character n-grams for $n = 2, 3, 4$ with and without stop words (6), word n-grams Jaccard coefficient for $n = 1, 2, 3, 4$ (4), word n-grams Jaccard coefficient without stop words for $n = 2, 4$ (2), word n-grams containment measure for $n = 1, 2$ (2) as well as pairwise word similarity (1).

TakeLab: We use several features described in Šarić et al. (2012)[12]: PathLen similarity (1 feature), corpus-based word similarity (3), vector space sentence similarity (1), n-gram overlap of tokens and lemmas for $n = 1, 2, 3$ (6), weighted word overlap for lowercased tokens and lemmas (2), normalized sentence length difference (1), shallow named entity features (4) and numbers overlap (3).

UMBC: We use several features described in Han et al. (2013): word n-gram similarity for $n = 1, 2, 3, 4$ (4 features). Moreover, we used word n-gram similarity for $n = 1$ where only nouns or only verbs where taken into account (2).

Readability Indicators: We use several features that are typically used as indicators for readability (Oelke et al., 2012): relative difference in sentence length, average word length in characters, number of nouns per sentence, number of verbs per sentence and noun-verb-ratio (5).

[8] Type-filtered subsets of A are denoted by A_{type}.

[9] See Sultan et al. (2014) for details on the formulae.

[10] Each content word is weighted by the similarity score achieved by word2vec for this type.

[11] Splitting $STS = T - P'$ into two features T and P' achieves better results than keeping it in the original form.

[12] takelab.fer.hr/sts/

5 STS Model

We compute STS scores using ν-SVR (Schölkopf et al., 2000) as implemented by LibSVM[13]. We use LibSVM's default SVR parameter settings without further optimization.

6 Interpretable STS Model

We align chunks using our word alignment (see Section 3). Because our word alignment itself does not rely on chunks, we extend its alignments using given chunk boundaries. If alignments overlap, we choose the longest alignment and discard the others. We do not differentiate between *SIMI* and *REL* – all *REL* alignments are considered as *SIMI* alignments.

For chunking we use the OpenNLP[14] chunker with the default model trained on CoNLL-2000 shared task data (Sang and Buchholz, 2000).

7 Results

For English we train on all available data sets from STS challenges in 2012 (Agirre et al., 2012), 2013 (Agirre et al., 2013) and 2014 (Agirre et al., 2014). For Spanish, each run trains on a different setting. Mean Pearson correlation is employed as an evaluation metric.

7.1 Subtask 2a – STS English

Table 1 presents the official scores of our system. Run *default* uses our system as it is. Run *themis* only relies on alignment features in the *belief* model, all other models are the same as for *default*. Our third run – *themisexp* – is identical to run *themis* except for one improvement: it penalizes scores of the *answers-students* dataset exponentially to cope with the high ratio of common content words that lead to over-estimation of similarity scores.

7.2 Subtask 2b – STS Spanish

Table 2 presents the official scores of our system. Run *trainEs* was trained on both Spanish test sets of 2014. Run *trainEn* was trained on all available English data sets. Run *trainMini* uses different training sets for each test set: *Wikipedia* model was trained on the 2014 *Wikipedia* test set and the *Newswire* model was trained on the *News* test set of 2014.

[13] www.csie.ntu.edu.tw/~cjlin/libsvm/

[14] opennlp.apache.org

Dataset	default	themis	themisexp
forum	0.6946 (10)	0.6946 (10)	0.6946 (10)
students	0.7505 (11)	0.7505 (11)	0.7784 (6)
belief	0.7521 (3)	0.7482 (6)	0.7482 (6)
headlines	0.8245 (7)	0.8245 (7)	0.8245 (7)
images	0.8527 (12)	0.8527 (12)	0.8527 (12)
Mean	0.7878 (8)	0.7873 (9)	**0.7942 (2)**

Table 1: Results (rank) of our three runs on English data.

Dataset	trainEs	trainMini	trainEn
Wikipedia	0.7055 (2)	0.7055 (1)	0.6763 (3)
Newswire	0.6830 (1)	0.6811 (2)	0.6705 (3)
Mean	**0.6905 (1)**	0.6893 (2)	0.6725 (3)

Table 2: Results (rank) of our three runs on Spanish data.

7.3 Subtask 2c – Interpretable STS

Our three runs only differ regarding the applied alignment scorer method: we use the *average* similarity score per alignment type as observed in STSint training data, the *most frequent* similarity score per alignment type as observed in STSint training data, and an STS *regression* model per alignment type trained on all available English STS data sets.

For subtrack *gold chunks*, our runs score 0.4885 to 0.4883 (F1 TYPE + SCORE) on headlines (ranks 10 - 12 out of 14) and 0.4296 to 0.4246 on images (ranks 8 - 10). Using *system chunks* we achieve scores of 0.4290 to 0.4284 on headlines (ranks 4–6 out of 10) and 0.3870 to 0.3806 on images (ranks 4–6).

8 Conclusions & Future Work

We presented our alignment-based STS system *ExB Themis*. Our system outperformed all other participants by a large margin on Spanish data. Furthermore, our system placed second on English data. *ExB Themis* proved to be the best multilingual STS system that easily can be adapted to further languages. We conclude that extensive feature extraction from word alignments is a very robust approach – especially when being applied to languages that lack high-quality resources.

In future work, we will investigate the influence of particular features in more detail and we want to enrich our model with structural information (Severyn et al., 2013; Sultan et al., 2014) and improved phrase similarity computation.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6 : A Pilot on Semantic Textual Similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 32–43, Atlanta, USA.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Ann Arbor, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, CO.

Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 435–440.

Aitor González-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the Sixth International Global WordNet Conference (GWC12)*, Matsue, Japan.

Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. UMBC EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*.

Christian Hänig, Stefan Bordag, and Stefan Thomas. 2014. Modular Classifier Ensemble Architecture for Named Entity Recognition on Low Resource Systems. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 113–116, Hildesheim, Germany.

Abhay Kashyap, Lushan Han, Roberto Yus, Jennifer Sleeman, Taneeya Satyapanich, Sunil Gandhi, and Tim Finin. 2014. Meerkat Mafia: Multilingual and Cross-Level Semantic Textual Similarity Systems. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 416–423, Dublin, Ireland.

Michael Lamar, Yariv Maron, Mark Johnson, and Elie Bienenstock. 2010. SVD and Clustering for Unsupervised POS Tagging. In *Proceedings of ACL 2010*, pages 215–219, Uppsala, Sweden.

Charles T. Meadow, Bert R. Boyce, and Donald H. Kraft. 1992. *Text Information Retrieval Systems*, volume 2. Academic Press San Diego.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the 21st national conference on Artificial intelligence*, pages 775–780.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*, pages 1–12, January.

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

Daniela Oelke, David Spretke, Andreas Stoffel, and Daniel A Keim. 2012. Visual readability analysis: How to make your writings easier to read. *IEEE Transactions on Visualization and Computer Graphics*, 18(5):662–674.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 Shared Task: Chunking. In *Proceedings of CoNLL 2000*, pages 127–132.

Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett. 2000. New Support Vector Algorithms. *Neural Computation*, 12(5):1207–1245.

Aliaksei Severyn, Massimo Nicosia, and Alessandro Moschitti. 2013. iKernels-Core: Tree Kernel Learning for Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 53–58, Atlanta, USA.

Robert Speer and Catherine Havasi. 2012. ConceptNet 5: A Large Semantic Network for Relational Knowledge. In *The Peoples Web Meets NLP*, pages 161–176.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. DLS@CU: Sentence Similarity from Word Alignment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 241–246, Dublin, Ireland.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. TakeLab: Systems for Measuring Semantic Text Similarity. In *First Joint Conference on Lexical and Computational Semantics*, pages 441–448, Montreal, Canada.