SOPA: Random Forests Regression for the Semantic Textual Similarity task

Davide Buscaldi, Jorge J. García Flores,
Laboratoire d'Informatique de Paris Nord, CNRS (UMR 7030)
Université Paris 13, Sorbonne Paris Cité, Villetaneuse, France
{buscaldi,jgflores}@lipn.univ-paris13.fr

Ivan V. Meza and Isaac Rodríguez
Instituto de Investigaciones en Matemticas Aplicadas y en Sistemas (IIMAS)
Universidad Nacional Autnoma de México (UNAM)
Ciudad Universitaria, DF, Mexico
ivanvladimir,isaac@turing.iimas.unam.mx

Abstract

This paper describes the system used by the LIPN-IIMAS team in the Task 2, Semantic Textual Similarity, at SemEval 2015, in both the English and Spanish sub-tasks. We included some features based on alignment measures and we tested different learning models, in particular Random Forests, which proved the best among those used in our participation.

1 Introduction

Our participation in SemEval 2015 was focused on solving the technical problems that afflicted our previous participation (Buscaldi et al., 2014) and including additional features based on alignments, such as the Sultan similarity (Sultan et al., 2014b) and the measure available in CMU Sphinx-4 (Lamere et al., 2003) for speech recognition. We baptised the new system SOPA from the Spanish word for "soup", since it uses a heterogeneous mix of features. Well aware of the importance that the training corpus and the regression algorithms have for the STS task, we used language models to select the most appropriate training corpus for a given text, and we explored some alternatives to the ν-Support Vector Regression (ν-SVR) (Schölkopf et al., 1999) used in our previous participations, specifically the Multi-Layer Perceptron (Bishop and others, 1995) and Random Forest (Breiman, 2001) regression algorithms. The obtained results show that Random Forests outperforms the other algorithms on every test set. We describe all the features in Section 2; the details on the learning algorithms and the training corpus selection process are described in Section 3, and the results obtained by the system are detailed in Section 4.

2 Similarity Measures

In this section we describe the measures used as features in our system. The total number of features used was 16 in English and 14 in Spanish. Since most measures have already been used in our previous participation, we provide only basic overview, referring the reader to the complete description in (Buscaldi et al., 2013) for further details. When POS tagging and NE recognition were required, we used the Stanford CoreNLP for English and Spanish (Manning et al., 2014).

2.1 WordNet-based Conceptual Similarity

This measure has been introduced in order to measure similarities between concepts with respect to an ontology. The similarity is calculated as follows: first of all, words in sentences p and q are lemmatised and mapped to the related WordNet synsets. All noun synsets are put into the set of synsets associated to the sentence, C_p and C_q, respectively. If the synsets are in one of the other POS categories (verb, adjective, adverb) we look for their derivationally related forms in order to find a related noun synset: if there exists one, we put this synset in C_p (or C_q). No disambiguation process is carried out, so we take all possible meanings into account.

Given C_p and C_q as the sets of concepts contained in sentences p and q, respectively, with $|C_p| \geq |C_q|$, the conceptual similarity between p and q is calcu-

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 132–137,
Denver, Colorado, June 4-5, 2015.

lated as:

$$ss(p,q) = \frac{\sum\limits_{c_1 \in C_p} \max\limits_{c_2 \in C_q} s(c_1, c_2)}{|C_p|}$$

where $s(c_1, c_2)$ is a conceptual similarity measure. Concept similarity can be calculated in different ways. We used a variation of the Wu-Palmer formula (Wu and Palmer, 1994) named "ProxiGenea3", introduced by (Dudognon et al., 2010), which is inspired by the analogy between a family tree and the concept hierarchy in WordNet. The ProxiGenea3 measure is defined as:

$$s(c_1, c_2) = \frac{1}{1 + d(c_1) + d(c_2) - 2 \cdot d(c_0)}$$

where c_0 is the most specific concept that is present both in the synset path of c_1 and c_2 (that is, the Least Common Subsumer or LCS). The function returning the depth of a concept is noted with d.

2.2 IC-based Similarity

This measure has been proposed by (Mihalcea et al., 2006) as a corpus-based measure which uses Resnik's Information Content (IC) and the Jiang-Conrath (Jiang and Conrath, 1997) similarity metric. This measure is more precise than the one introduced in the previous subsection because it takes into account also the importance of concepts and not only their relative position in the hierarchy. We refer to (Buscaldi et al., 2013) and (Mihalcea et al., 2006) for a detailed description of the measure. The idf weights for the words were calculated using the Google Web 1T (Brants and Franz, 2006) frequency counts, while the IC values used are those calculated by Ted Pedersen (Pedersen et al., 2004) on the British National Corpus[1].

2.3 Syntactic Dependencies

This measure tries to capture the syntactic similarity between two sentences using dependencies. Previous experiments showed that converting constituents to dependencies still achieved best results on out-of-domain texts (Le Roux et al., 2012), so we decided to use a 2-step architecture to obtain syntactic dependencies. First we parsed pairs of sentences with the LORG parser[2]. Second we converted the resulting parse trees to Stanford dependencies.

Given the sets of parsed dependencies D_p and D_q, for sentence p and q, a dependency $d \in D_x$ is a triple (l, h, t) where l is the dependency label (for instance, *dobj* or *prep*), h the governor and t the dependant. The similarity measure between two syntactic dependencies $d_1 = (l_1, h_1, t_1)$ and $d_2 = (l_2, h_2, t_2)$ is the levenshtein distance between the labels l_1 and l_2 multiplied by the average of $idf_h * s(h_1, h_2)$ and $idf_t * s(t_1, t_2)$, where idf_h and idf_t are the inverse document frequencies calculated on Google Web 1T for the governors and the dependants (we retain the maximum for each pair), respectively, and s is the ProxiGenea3 measure. NOTE: This measure was used only in the English sub-task.

2.4 Information Retrieval-based Similarity

Let us consider two texts p and q, an IR system S and a document collection D indexed by S. This measure is based on the assumption that p and q are similar if the documents retrieved by S for the two texts, used as input queries, are ranked similarly.

Let be $L_p = \{d_{p_1}, \ldots, d_{p_K}\}$ and $L_q = \{d_{q_1}, \ldots, d_{q_K}\}$, $d_{x_i} \in D$ the sets of the top K documents retrieved by S for texts p and q, respectively. Let us define $s_p(d)$ and $s_q(d)$ the scores assigned by S to a document d for the query p and q, respectively. Then, the similarity score is calculated as:

$$sim_{IR}(p,q) = 1 - \frac{\sum\limits_{d \in L_p \cap L_q} \frac{\sqrt{(s_p(d) - s_q(d))^2}}{\max(s_p(d), s_q(d))}}{|L_p \cap L_q|}$$

if $|L_p \cap L_q| \neq \emptyset$, 0 otherwise.

For the participation in the English sub-task we indexed a collection composed by the AQUAINT-2[3] and the English NTCIR-8[4] document collections, using the Lucene[5] 4.2 search engine with BM25 similarity. We indexed also DBPedia[6] abstracts and the UkWaC (Ferraresi et al., 2008), but they were used to produce two additional features (separate

[1] http://www.d.umn.edu/ tpederse/similarity.html

[2] https://github.com/CNGLdlab/LORG-Release

[3] http://www.nist.gov/tac/data/data_desc.html#AQUAINT-2

[4] http://metadata.berkeley.edu/NTCIR-GeoTime/ntcir-8-databases.php

[5] http://lucene.apache.org/core

[6] http://www.dbpedia.org/

from the basic IR one). The Spanish index was created using the Spanish QA@CLEF 2005 (agencia EFE1994-95, El Mundo 1994-95) and multiUN (Eisele and Chen, 2010) collections. The K value was set to 70 after a study detailed in (Buscaldi, 2013). Another IR-based feature was derived by the rank-biased overlap measure introduced by (Webber et al., 2010) which compares rankings without the need of weights. In total, we had 4 IR-based measures for English and 2 for Spanish.

2.5 N-gram Based Similarity

This measure tries to capture the fact that similar sentences have similar n-grams, even if they are not placed in the same positions. The measure is based on the Clustered Keywords Positional Distance (CKPD) model proposed in (Buscaldi et al., 2009) for the passage retrieval task.

The similarity between a text fragment p and another text fragment q is calculated as:

$$sim_{ngrams}(p,q) = \sum_{\forall x \in Q} \frac{h(x,P)}{\sum_{i=1}^{n} w_i d(x, x_{max})}$$

Where P is the set of the heaviest *n*-grams in p where all terms are also contained in q; Q is the set of all the possible n-grams in q, and n is the total number of terms in the longest sentence. The weights for each term w_i are calculated as $w_i = 1 - \frac{log(n_i)}{1+log(N)}$ where n_i is the frequency of term t_i in the Google Web 1T collection, and N is the frequency of the most frequent term in the Google Web 1T collection. The weight for each n-gram ($h(x,P)$), with $|P| = j$ is calculated as:

$$h(x,P) = \begin{cases} \sum_{k=1}^{j} w_k & \text{if } x \in P \\ 0 & \text{otherwise} \end{cases}$$

The function $d(x, x_{max})$ determines the minimum distance between a *n*-gram x and the heaviest one x_{max} as the number of words between them.

2.6 Geographical Context Similarity

This measure tries to measure if the two sentences refer to events that took place in the same geographical area. It is based on the observation that the compatibility of the geographic context between the sentences is an important clue to determine whether the sentences are related or not, especially in news. We built a database of geographically-related entities, using geo-WordNet (Buscaldi and Rosso, 2008) and expanding it with all the synsets that are related to a geographically grounded synset. This implies that also adjectives and verbs may be used as clues for the identification of the geographical context of a sentence. For instance, "Afghan" is associated to "Afghanistan", "Sovietize" to "Soviet Union", etc. The Named Entities of type PER (Person) are also used as clues: we use Yago[7] to check whether the NE corresponds to a famous leader or not, and in the affirmative case we include the related nation to the geographical context of the sentence. For instance, "Merkel" is mapped to "Germany". Given G_p and G_q the sets of places found in sentences p and q, respectively, the geographical context similarity is calculated as follows:

$$sim_{geo}(p,q) = 1 - \log_K \left(1 + \frac{\sum_{x \in G_p} \min_{y \in G_q} d(x,y)}{\max(|G_p|, |G_q|)} \right)$$

Where $d(x,y)$ is the spherical distance in Km. between x and y, and K is a normalization factor set to 10000 Km. to obtain similarity values between 1 and 0. If no toponyms or geographically groundable entities are found in either sentences, then the geographic context similarity is set to 1.

2.7 Word Alignment Similarity

This similarity metric is based on the work of (Sultan et al., 2014b; Sultan et al., 2014a). The metric calculates a similarity score based on an alignment between two texts. It starts with an alignment between similar words, it proceeds to align similar name entities, to continue with words with similar content, to finally align stop words. In the case of content words, it proposes to use the syntactic context to identify similar words. At the end, the similarity is calculated as a harmonic mean between the ratios of align words from sentence one to sentence two, and from sentence two to sentence one.

CMU Sphinx-4 (Lamere et al., 2003) is a speech recognition system that includes an alignment function that is used to align speech transcriptions with

[7] http://www.mpi-inf.mpg.de/yago-naga/yago/

text. We took one of the sentence as a reference and the other one as a transcription and we used the output of the Sphinx alignment measure as a feature.

2.8 Other Measures

In addition to the above text similarity measures, we used also the difference in size between sentences and the following measures:

Cosine

Cosine distance calculated between $\mathbf{p} = (w_{p_1}, \ldots, w_{p_n})$ and $\mathbf{q} = (w_{q_1}, \ldots, w_{q_n})$, the vectors of $tf.idf$ weights associated to sentences p and q, with idf values calculated on Google Web 1T.

Edit Distance

This similarity measure is calculated using the Levenshtein distance on characters between the two sentences.

Named Entity Overlap

This is a per-class overlap measure (in this way, "France" as an Organization does not match "France" as a Location) calculated using the Dice coefficient between the sets of NEs found, respectively, in sentences p and q.

Skip-gram Similarity

This measure is obtained as the dice coefficient calculated between the set of skip-grams contained in the two sentences.

3 Learning Models

Besides the ν-Support Vector Regression (ν-SVR) (Schölkopf et al., 1999) used in previous participation, we used Multilayer Perceptron and Random Forests. The Multilayer perceptron (Bishop and others, 1995) is a neural network model which has several interesting properties, such as robustness and nonlinearity. Our implementation uses a simple gradient descent learning algorithm with backpropagation and one hidden layer with 5 units. Random Forests (Breiman, 2001) are an ensemble learning method based on boosting and bagging of classification trees. In our experiments, we used Random Forests with 10 bootstrap samples.

In our runs, we selected a subset of the training set according to a similarity measure between the test and the training set based on a 1- to 3-grams language model and average sentence length. The idea behind this selection process is that learning sentence similarities on a specific type of text will increase the accuracy of predictions on text with similar characteristics: image descriptions are usually written in a very different form than word definitions or forum answers. For each coherent subset of the training set, we built a language model $L_m = (G_1, G_2, G_3)$ where G_n is the distribution frequency of n-grams in the subset. We obtained the same for the input dataset (L_i) and we calculated $S(L_m, L_i) = (b(L_{m1}, L_{i1}) + 2 * b(L_{m2}, L_{i2}) + 3 * b(L_{m3}, L_{i3}))/6$ where $b(F_1, F_2)$ is the Bhatthacharyya distance between the distributions F_1 and F_2. We selected only those training dataset where $S(L_m, L_i) > 0.2$. In Table 3 we show the comparison of the results obtained with such selection (the official ones) and those obtained using the complete training set (not submitted). The complete English training set was composed by the data from SemEval STS 2012, 2013 and 2014. In Spanish, we used our 2014 training set, which included the automatically translated English 2012-2013 pairs from STS and a corpus we made from RAE[8] definitions, and the 2014 Spanish test set.

4 Results

Table 1 and 2 presents our results our runs in SemEval 2015 (Agirre et al., 2015). Our participation consisted on three runs for three different machine learning approaches to regressions: Support Vector Regression (*LIPN-SVM*), Multi Layer Perceptron (*LIPN-MLP*) and Random Forest (*LIPN-RF*). The *LIPN-RF* configuration was our best one and was ranked 25th run-wise and 14th system wise for the English corpora; 5th run-wise and 3rd system-wise for Spanish. Our English system had better overall performance than Spanish. The best performance was reached for the *Believe* dataset in English and *News* dataset in Spanish.

Part of our proposal uses a language model to select a subset of the corpus used to train the regression. Table 3 shows performance with the full dataset and the selected training corpus for the En-

[8] "Real Academia Española de la lengua" Spanish dictionary: `http://www.rae.es`

	Answer-forums	Answer-students	Headlines	Believe	Images	Overall
LIPN-RF	**0,6709**	**0,5914**	**0,7243**	**0,8123**	**0,8414**	**0,7356**
LIPN-MLP	0,6178	0,5864	0,6886	0,8121	0,8184	0,7175
LIPN-SVM	0,5918	0,5718	0,7028	0,7985	0,8104	0,7070

Table 1: English results (Official runs).

	Wikipedia	News	Overall
LIPN-RF	**0,5637**	**0,5655**	**0,5649**
LIPN-MLP	0,25257	0,5342	0,4401
LIPN-SVM	0,4194	0,4007	0,4069

Table 2: Spanish results (Official runs).

glish dataset with the three regression approaches. The overall score points that the corpus selection was not beneficial. The most significant improvement was concentrated on the *Answer-students* dataset, in this case the performance felt 0,0588 points.

We checked the contribution of each feature using the relief attribute selection measure (Kononenko, 1994) over the English training set. The best feature was the WordNet one, followed by Sultan and IC-based similarity. The worst features were Rank-biased Overlap followed by NE Overlap and the Geographic context similarity (however, apart from RBO, the other ones don't have complete coverage). The other features have a statistically similar contribution.

5 Conclusions and Future Work

The new learning models adopted were particularly effective, outperforming the Support Vector Regression algorithm that we used in our previous participation. The alignment measure based on Sultan was also very effective, as indicated by feature selection. On the other hand, our corpus selection strategy did not prove useful in general, although it provided slight improvements on specific corpora. We will need to further analyse these results to understand how SOPA can still be improved.

Acknowledgements

This work is supported/ partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the progam "Investissements d'Avenir" (reference: ANR-10-LABX-0083).

References

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, June.

Christopher M Bishop et al. 1995. Neural networks for pattern recognition.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram corpus version 1.1.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Davide Buscaldi and Paolo Rosso. 2008. Geo-WordNet: Automatic Georeferencing of WordNet. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco.

Davide Buscaldi, Paolo Rosso, José Manuel Gómez, and Emilio Sanchis. 2009. Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems (JIIS)*, 34(2):113–134.

Davide Buscaldi, Joseph Le Roux, Jorge J. Garcia Flores, and Adrian Popescu. 2013. LIPN-CORE: Semantic Text Similarity using n-grams, WordNet, Syntactic Analysis, ESA and Information Retrieval based Features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 162–168, Atlanta, Georgia, USA, June.

Davide Buscaldi, Jorge J García Flores, Joseph Le Roux, Nadi Tomeh, and Belem Priego-Sanchez. 2014. LIPN: Introducing a new Geographical Context Similarity Measure and a Statistical Similarity Measure based on the Bhattacharyya coefficient. In *SemEval 2014*, pages 400–405.

Davide Buscaldi. 2013. Une mesure de similarité sémantique basée sur la Recherche d'Information. In

	Answer-forums	Answer-students	Headlines	Believe	Images	Overall
Selected						
LIPN-RF	**0,6709**	0,5914	0,7243	0,8123	**0,8414**	0,7244
LIPN-MLP	0,6178	0,5864	0,6886	0,8121	0,8184	0,6986
LIPN-SVM	0,5918	0,5718	0,7028	0,7985	0,8104	0,6894
Full						
LIPN-RF	0,6418	**0,6502**	**0,7320**	**0,8155**	**0,8301**	**0,7339**
LIPN-MLP	0,6252	0,6213	0,8047	0,6856	0,8047	0,7120
LIPN-SVM	0,5701	0,6177	0,7939	0,7003	0,7939	0,7001

Table 3: Comparison of the results obtained with corpus selection and using the full corpus.

5ème Atelier Recherche d'Information SEmantique - RISE 2013, pages 81–91, Lille, France, July.

Damien Dudognon, Gilles Hubert, and Bachelin Jhonn Victorino Ralalason. 2010. Proxigénéa : Une mesure de similarité conceptuelle. In *Proceedings of the Colloque Veille Stratégique Scientifique et Technologique (VSST 2010)*.

Andreas Eisele and Yu Chen. 2010. MultiUN: A Multilingual Corpus from United Nation Documents. In Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaard, Khalid Choukri, and Nicoletta Calzolari (Conference Chair), editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872, 5.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating UkWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pages 19–33.

Igor Kononenko. 1994. Estimating attributes: analysis and extensions of RELIEF. In *Machine Learning: ECML-94*, pages 171–182.

Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. 2003. The cmu sphinx-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, pages 2–5. Citeseer.

Joseph Le Roux, Jennifer Foster, Joachim Wagner, Rasul Samad Zadeh Kaljahi, and Anton Bryl. 2012. DCU-Paris13 Systems for the SANCL 2012 Shared Task. In *The NAACL 2012 First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, pages 1–4, Montréal, Canada.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

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 - Volume 1*, AAAI'06, pages 775–780.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, pages 38–41, Stroudsburg, PA, USA.

Bernhard Schölkopf, Peter Bartlett, Alex Smola, and Robert Williamson. 1999. Shrinking the tube: a new support vector regression algorithm. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 330–336, Cambridge, MA, USA.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014a. Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014b. DLS@ CU: Sentence Similarity from Word Alignment. *SemEval 2014*, page 241.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):20.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA.