LIPN: Introducing a new Geographical Context Similarity Measure and a Statistical Similarity Measure Based on the Bhattacharyya Coefficient

Davide Buscaldi, Jorge J. García Flores, Joseph Le Roux, Nadi Tomeh
Laboratoire d'Informatique de Paris Nord, CNRS (UMR 7030)
Université Paris 13, Sorbonne Paris Cité, Villetaneuse, France
{buscaldi,jgflores,joseph.le-roux,nadi.tomeh}@lipn.univ-paris13.fr

Belém Priego Sanchez
Laboratoire LDI (Lexique, Dictionnaires, Informatique)
Université Paris 13, Sorbonne Paris Cité, Villetaneuse, France
LKE, FCC, BUAP, San Manuel, Puebla, Mexico
belemps@gmail.com

Abstract

This paper describes the system used by the LIPN team in the task 10, Multilingual Semantic Textual Similarity, at SemEval 2014, in both the English and Spanish sub-tasks. The system uses a support vector regression model, combining different text similarity measures as features. With respect to our 2013 participation, we included a new feature to take into account the geographical context and a new semantic distance based on the Bhattacharyya distance calculated on co-occurrence distributions derived from the Spanish Google Books n-grams dataset.

1 Introduction

After our participation at SemEval 2013 with LIPN-CORE (Buscaldi et al., 2013) we found that geography has an important role in discriminating the semantic similarity of sentences (especially in the case of newswire). If two events happened in a different location, their semantic relatedness is usually low, no matter if the events are the same. Therefore, we worked on a similarity measure able to capture the similarity between the geographic contexts of two sentences. We tried also to reinforce the semantic similarity features by introducing a new measure that calculates word similarities on co-occurrence distributions extracted from Google Books bigrams. This measure was introduced only for the Spanish runs, due to time constraints. The regression model used to integrate the features was the ν-Support Vector Regression model (ν-SVR) (Schölkopf et al., 1999) implementation provided by LIBSVM (Chang and Lin, 2011), with a radial basis function kernel with the standard parameters ($\nu = 0.5$). We describe all the measures in Section 2; the results obtained by the system are detailed in Section 3.

2 Similarity Measures

In this section we describe the measures used as features in our system. The description of measures already used in our 2013 participation is less detailed than the description of the new ones. Additional details on the measures may be found in (Buscaldi et al., 2013). When POS tagging and NE recognition were required, we used the Stanford CoreNLP[1] for English and FreeLing[2] 3.1 for Spanish.

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

[1] http://www-nlp.stanford.edu/software/corenlp.shtml

[2] http://nlp.lsi.upc.edu/freeling/

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 400–405,
Dublin, Ireland, August 23-24, 2014.

$|C_p| \geq |C_q|$, the conceptual similarity between p and q is calculated 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[3].

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[4]. Second we converted the resulting parse trees to Stanford dependencies[5].

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 $idf_h * s_{WN}(h_1, h_2)$ and $idf_t * s_{WN}(t_1, t_2)$, 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_{WN} is calculated using formula **??**. 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[6] and the English NTCIR-8[7] document collections, using the Lucene[8] 4.2 search engine with BM25 similarity. The Spanish index was created using the Spanish QA@CLEF 2005 (agencia EFE1994-95, El Mundo 1994-95) and multiUN

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

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

[5] We used the default built-in converter provided with the Stanford Parser (2012-11-12 revision).

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

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

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

(Eisele and Chen, 2010) collections. The K value was set to 70 after a study detailed in (Buscaldi, 2013).

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

We observed that in many sentences, especially those extracted from news corpora, the compatibility of the geographic context between the sentences is an important clue to determine if the sentences are related or not. This measure tries to measure if the two sentences refer to events that took place in the same geographical area. 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[9] 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.

2.7 2-grams “Spectral” Distance

This measure is used to calculate the semantic similarity of two words on the basis of their context, according to the distributional hypothesis. The measure exploits bi-grams in the Google Books n-gram collection[10] and is based on the distributional hypothesis, that is, “words that tend to appear in similar contexts are supposed to have similar meanings”. Given a word w, we calculate the probability of observing a word x knowing that it is preceded by w as $p(x|w) = p(w \cap x)/p(w) = c(\text{“}wx\text{”})/c(\text{“}w\text{”})$, where $c(\text{“}wx\text{”})$ is the number of bigrams “w x” observed in Google Books (counting all publication years) 2-grams and $c(\text{“}w\text{”})$ is the number of occurrences of w observed in Google Books 1-grams. We calculate also the probability of observing a word y knowing that it is followed by w as $p(y|w) = p(w \cap y)/p(w) = c(\text{“}yw\text{”})/c(\text{“}w\text{”})$. In such a way, we may obtain for a word w_i two probability distributions $D_p^{w_i}$ and $D_f^{w_i}$ that can be compared to the distributions obtained in the same way for another word w_j. Therefore, we calculate the distance of two words comparing the distribution probabilities built in this way, using the Bhattacharyya coefficient:

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

[10] https://books.google.com/ngrams/datasets

$$s_f(w_i, w_j) = -\log\left(\sum_{x \in X} \sqrt{D_f^{w_i}(x) * D_f^{w_j}(x)}\right)$$

$$s_p(w_i, w_j) = -\log\left(\sum_{x \in X} \sqrt{D_p^{w_i}(x) * D_p^{w_j}(x)}\right)$$

the resulting distance between w_i and w_j is calculated as the average between $s_f(w_i, w_j)$ and $s_p(w_i, w_j)$. All words in sentence p are compared to the words of sentence q using this similarity value. The words that are semantically closer are paired; if a word cannot be paired (average distance with any of the words in the other sentence > 10), then it is left unpaired. The value used as the final feature is the averaged sum of all distance scores.

2.8 Other Measures

In addition to the above text similarity measures, we used also the following common 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.

3 Results

3.1 Spanish

In order to train the Spanish model, we translated automatically all the sentences in the English SemEval 2012 and 2013 using Google Translate. We also built a corpus manually using definitions from the RAE[11] (Real Academia Española de la Lengua). The definitions were randomly extracted and paired at different similarity levels (taking into account the Dice coefficient calculated on the definitions bag-of-words). Three annotators gave independently their similarity judgments on these paired definitions. A total of 200 definitions were annotated for training. The official results for the Spanish task are shown in Table 1. In Figure 1 we show the results obtained by taking into account each individual feature as a measure of similarity between texts. These results show that the combination was always better than the single features (as expected), and the feature best able to capture semantic similarity alone was the cosine distance. In Table 2 we show the results of the ablation test, which shows that the features that most contributed to improve the results were the IR-based similarity for the news dataset and the cosine distance for the Wikipedia dataset. The worst feature was the NER overlap (not taking into account it would have allowed us to gain 2 places in the final rankings).

	Wikipedia	**News**	**Overall**
LIPN-run1	0.65194	0.82554	0.75558
LIPN-run2	**0.71647**	**0.8316**	**0.7852**
LIPN-run3	0.71618	0.80857	0.77134

Table 1: Spanish results (Official runs).

The differences between the three submitted runs are only in the training set used. `LIPN-run1` uses all the training data available together, `LIPN-run3` uses a training set composed by the translated news for the news dataset and the RAE training set for the Wikipedia dataset; finally, the best run `LIPN-run2` uses the same training sets of run3 together to build a single model.

3.2 English

Our participation in the English task was hampered by some technical problems which did not allow us to complete the parsing of the tweet data in time. As a consequence of this and some errors in the scripts launched to finalize the experiments, the submitted results were incomplete and we were able to detect the problem only after the submission. We show in Table 3 the official results of run1 with the addition of the results on the OnWN dataset calculated after the participation to the task.

[11] http://www.rae.es/

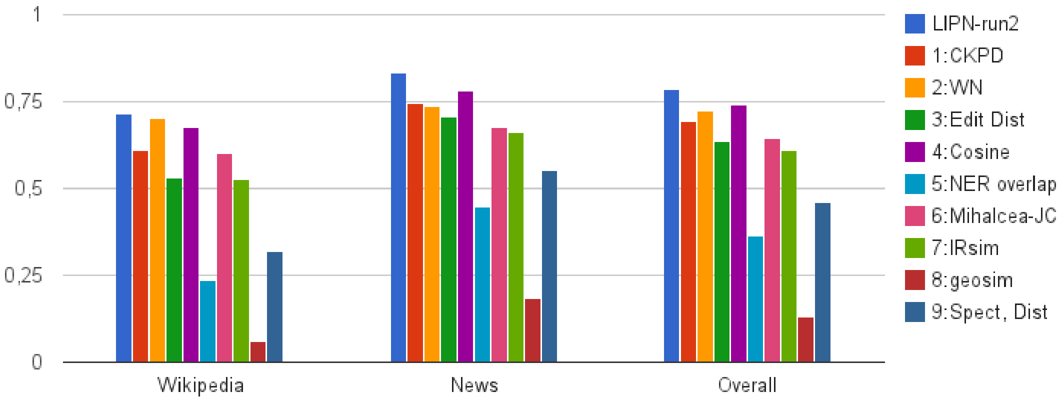

Figure 1: Spanish task: results taking into account the individual features as semantic similarity measures.

Ablated feature	**Wikipedia**	**News**	**Overall**	**diff**
LIPN-run2 (none)	0.7165	0.8316	0.7852	0.00%
1:CKPD	0.7216	0.8318	0.7874	0.22%
2:WN	0.7066	0.8277	0.7789	−0.63%
3:Edit Dist	0.708	0.8242	0.7774	−0.78%
4:Cosine	**0.6849**	0.8235	0.7677	**−1.75%**
5:NER overlap	**0.7338**	**0.8341**	0.7937	**0.85%**
6:Mihalcea-JC	0.7103	0.8301	0.7818	−0.34%
7:IRsim	0.7161	**0.8026**	0.7677	**−1.74%**
8:geosim	0.7185	0.8325	0.7865	0.14%
9:Spect. Dist	0.7243	0.8311	0.7880	0.28%

Table 2: Spanish task: ablation test.

Dataset	**Correlation**
Complete (official + OnWN)	0.6687
Complete (only official)	0.5083
deft-forum	0.4544
deft-news	0.6402
headlines	0.6527
images	0.8094
OnWN (unofficial)	0.8039
tweet-news	0.5507

Table 3: English results (Official run + unofficial OnWN).

4 Conclusions and Future Work

The introduced measures were studied on the Spanish subtask, observing a limited contribution from geographic context similarity and spectral distance. The IR-based measure introduced in 2013 proved to be an important feature for newswire-based datasets as in the 2013 English task, even when trained on a training set derived from automatic translation, which include many errors. Our participation in the English subtask was inconclusive due to the technical faults experienced to produce our results. We will nevertheless take into account the lessons learned in this participation for future ones.

Acknowledgements

Part of this work has been carried out with the support of LabEx-EFL (Empirical Foundation of Linguistics) strand 5 (computational semantic analysis). We are also grateful to CoNACyT (Consejo NAcional de Ciencia y Tecnologia) for support to

this work.

References

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

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. Association for Computational Linguistics.

Davide Buscaldi. 2013. Une mesure de similarité sémantique basée sur la recherche d'information. In *5ème Atelier Recherche d'Information SEmantique - RISE 2013*, pages 81–91, Lille, France, July.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

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. European Language Resources Association (ELRA), 5.

J.J. Jiang and D.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.

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, June.

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. AAAI Press.

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. Association for Computational Linguistics.

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. MIT Press.

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. Association for Computational Linguistics.