HulTech: A General Purpose System for Cross-Level Semantic Similarity based on Anchor Web Counts

Jose G. Moreno **Rumen Moraliyski** **Asma Berrezoug** **Gaël Dias**
Normandie University
UNICAEN, GREYC CNRS
F-14032 Caen, France
firstname.lastname@unicaen.fr

Abstract

This paper describes the HULTECH team participation in Task 3 of SemEval-2014. Four different subtasks are provided to the participants, who are asked to determine the semantic similarity of cross-level test pairs: paragraph-to-sentence, sentence-to-phrase, phrase-to-word and word-to-sense. Our system adopts a unified strategy (general purpose system) to calculate similarity across all subtasks based on word Web frequencies. For that purpose, we define ClueWeb InfoSimba, a cross-level similarity corpus-based metric. Results show that our strategy overcomes the proposed baselines and achieves adequate to moderate results when compared to other systems.

1 Introduction

Similarity between text documents is considered a challenging task. Recently, many works concentrate on the study of semantic similarity for multi-level text documents (Pilehvar et al., 2013), but skipping the cross-level similarity task. In the later, the underlying idea is that text similarity can be considered between pairs of text documents at different granularities levels: paragraph, sentence, phrase or word. One obvious particularity of this task is that text pairs may not share the same characteristics of size, context or structure, i.e., the granularity level.

In task 3 of SemEval-2014, two different strategies have been proposed to solve this issue. On the one hand, participants may propose a combination of individual systems, each one solving a particular subtask. On the other hand, a general purpose system may be proposed, which deals with all the subtasks following the exact same strategy.

In this paper, we describe a language-independent corpus-based general purpose system, which relies on a huge freely available Web collection called Anchor-ClueWeb12 (Hiemstra and Hauff, 2010). In particular, we calculate ClueWeb InfoSimba[1] a cross-level semantic similarity based on word-word frequencies. Indeed, these frequencies are captured by the use of a collocation metric called SCP[2] (Silva et al., 1999), which has similar properties as the well studied PMI-IR (Turney, 2001) but does not over-evaluate rare events.

Our system outputs a normalized (between 0 and 1) similarity value between two pieces of texts. However, the subtasks proposed in task 3 of SemEval-2014 include a different scoring scale between 0 and 4. To solve this issue, we applied linear, polynomial and exponential regressions as three different runs. Results show that our strategy overcomes the proposed baselines and achieves adequate to moderate results when compared to other systems.

2 System Description

Our system is based on a reduced version of the ClueWeb12 dataset called Anchor ClueWeb12 and an informative attributional similarity measure called InfoSimba (Dias et al., 2007) adapted to this dataset.

2.1 Anchor ClueWeb12 Dataset

The Anchor ClueWeb12 dataset contains 0.5 billion Web pages, which cover about 64% of the total number of Web pages in ClueWeb12. The particularity of Anchor ClueWeb12 is that each Web page is represented by the anchor texts of the links pointing to it in ClueWeb12. Web pages are indexed not on their content but on their references. As such, the size of the index is drastically reduced and the overall results are consistent with full text indexing as discussed in (Hiemstra and Hauff, 2010).

For development purposes, this dataset was indexed in Solr 4.4 on a desktop computer using a batch indexing script. Particularly, each compressed part file of the Anchor ClueWeb12 was uncompressed, preprocessed and indexed in a sequential way using the features of incremental indexing offered by Solr (Smiley and Pugh, 2009).

2.2 InfoSimba

In (Dias et al., 2007), the authors proposed the hypothesis that two texts are similar if they share related (eventually different) constituents. So, their concept of simi-

[1] It is a Web version of InfoSimba (Dias et al., 2007).

[2] Symmetric Conditional Probability.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 305–308,
Dublin, Ireland, August 23-24, 2014.

larity is not any more based on the exact match of constituents but relies on related constituents (e.g. words). For example, it is clear that the following text pieces extracted from the sentence-to-phrase subtask are related[3] although they do not share any word.

1. *he is a nose-picker*

2. *an uncouth young man*

The InfoSimba similarity measure models this phenomenon evaluating individual similarities between all possible words pairs. Indeed, each piece of text is represented by the vector of its words. So, given two pieces of texts X_i and X_j, their similarity is defined in Equation 1 where $SCP(.,.)$ is the Symmetric Conditional Probability association measure proposed in (Silva et al., 1999) and defined in Equation 2.

$$IS(X_i, X_j) = \frac{1}{pq} \sum_{k=1}^{p} \sum_{l=1}^{q} SCP(w_{ik}, w_{jl}). \quad (1)$$

$$SCP(w_{ik}, w_{jl}) = \frac{P(w_{ik}, w_{jl})^2}{P(w_{ik}) \times P(w_{jl})}. \quad (2)$$

Following the previous example, the InfoSimba value between the two vectors $X_1 = \{\text{“}he\text{”}, \text{“}is\text{”}, \text{“}a\text{”}, \text{“}nose\text{-}picker\text{”}\}$ and $X_2 = \{\text{“}an\text{”}, \text{“}uncouth\text{”}, \text{“}young\text{”}, \text{“}man\text{”}\}$ is an average weight formed by all possible words pairs associations as illustrated in Figure 1. Note that each vertex is a word of a X_l vector and each edge is weighted by the $SCP(.,.)$ value of the connected words. In particular, each w_{ij} corresponds to the word at the j^{th} position in vector X_i, $P(.,.)$ is the joint probability of two words appearing in the same document, $P(.)$ is the marginal probability of any word appearing in a document and p (resp. q) is the size of the vector X_i (resp. X_j).

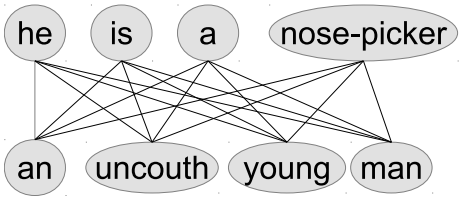

Figure 1: Pairs of words evaluated when InfoSimba is calculated.

In the case of task 3 of SemEval-2014, each text pair is represented by two word vectors for which a modified version of InfoSimba, ClueWeb InfoSimba, is computed.

[3]The score of this pair (#85) in the training set is the maximum value 4.

2.3 ClueWeb InfoSimba

The final similarity metric, called ClueWeb InfoSimba ($CWIS$), between two pieces of texts is defined in Equation 3, where $hits(w)$ returns the number of documents retrieved by Solr over Anchor ClueWeb12 for the query w and $hits(w_a \wedge w_b)$ is the number of documents retrieved when both words are present simultaneously. In this case, SCP is modified into SCP-IR similarly as PMI is to PMI-IR, i.e., using hits counts instead of probability values (see Equation 4).

$$CWIS(X_i, X_j) = \frac{1}{pq} \sum_{k=1}^{p} \sum_{l=1}^{q} SCP - IR(w_{ik}, w_{jl}). \quad (3)$$

$$SCP - IR(w_{ik}, w_{jl}) = \frac{hits(w_{ik} \wedge w_{jl})^2}{hits(w_{ik}).hits(w_{jl})}. \quad (4)$$

2.4 System Input

The task 3 of SemEval-2014 consists of (1) paragraph-to-sentence, (2) sentence-to-phrase, (3) phrase-to-word and (4) word-to-sense subtasks. Before submitting the pieces of texts to our system, we first performed simple stop-words removal with the NLTK toolkit (Bird et al., 2009). Note that in the case of the word-to-sense subtask, the similarity is performed over the word itself and the gloss of the corresponding sense[4].

2.5 Output Values Transformations

The $CWIS(.,.)$ similarity metric returns a value between 0 and 1. However, the subtasks suppose that each pair must be attributed a score between 0 and 4. As such, an adequate scale transformation must be performed. For that purpose, we proposed linear, polynomial and exponential regressions and submitted three different runs, one for each regression[5]. Note that the regressions have been tuned on the training dataset using the respective R regression functions with default parameters:

- $lm(y \sim x)$,
- $lm(y \sim x + I(x^2) + I(x^3))$,
- $lm(log(y + \epsilon) \sim x)$,

where ϵ[6] is a small value included to avoid undefined log values. The regression results on the test datasets are presented in Figure 2.

[4]Glosses are obtained from WordNet using the sense id provided for the task by the organizers.

[5]In the case of linear and exponential, these are monothetic functions therefore ranking-based evaluation metrics give the same score before and after this step.

[6]In our experiments, this value was set to 0.001.

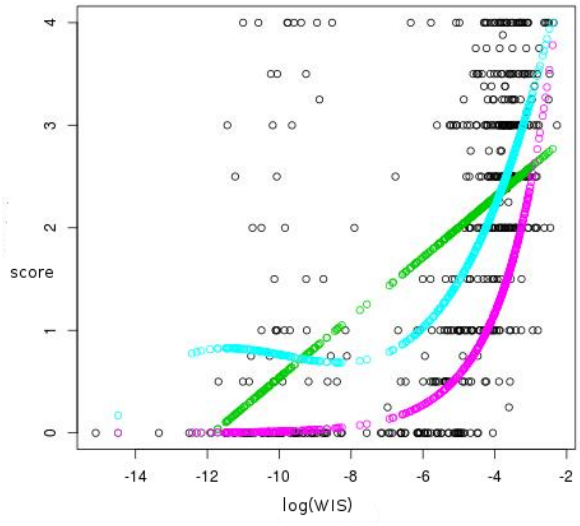

Figure 2: Linear, polynomial and exponential predictions for the test dataset of the paragraph-to-sentence subtask (colored dots). Black dots correspond to the obtained ClueWeb InfoSimba value versus the manually assigned score in the training dataset.

3 Evaluation and Results

For evaluation purposes, two metrics have been selected by the organizers: Pearson correlation (Pearson, 1895) and Spearman's rank correlation (Hollander and Wolfe, 1973). Detailed information about the evaluation setup can be found in the task discussion paper (Jurgens et al., 2014).

All results are given in Tables 1 and 2 for each run. Note that the baseline metric is calculated for the longest common string (LCS) and that each regression has been tuned on the training dataset for each one of the four tasks.

First, in almost all cases, the results outperform the baseline. Second, performances show that with a certain amount of information (longer pieces of texts), interesting results can be obtained. However, when the size decreases, the performance diminishes and extra information is certainly needed to better capture the semantics between two pieces of text. Third, the polynomial regression provides better results for the Pearson correlation evaluation, while for the Rho test, linear and polynomial regressions get the lead. Note that this situation depends on the data distribution and cannot be seen as a conclusive remark. However, it is certainly an important subject of study for our unsupervised methodology.

Another key point is that training examples were used only for evaluation purposes[7]. In the case of Spearman's rank correlation, the linear and exponential transformations obviously show exact same values (See Table 2).

[7]For Pearson correlation, valid interval was fixed to [0,4].

4 Conclusions

In this paper, we proposed a general purpose system to deal with cross-level text similarity. The aim of our research was to push as far as possible the limits of language-independent corpus-based solutions in a general context of text similarity. We were also concerned with reproducibility and as such we exclusively used publicly available datasets and tools[8]. The results clearly show the limits of a simple solution based on word statistics. Nevertheless, the framework can easily be empowered with the straightforward introduction of more competitive resources.

Acknowledgement

The authors would like to thank the University of Mostaganem (Algeria) for providing an internship to Asma Berrezoug at the Normandie University.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Gaël Dias, Elsa Alves, and José Gabriel Pereira Lopes. 2007. Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation. In *Proceedings of AAAI*, pages 1334–1339.

Djoerd Hiemstra and Claudia Hauff. 2010. Mirex: Mapreduce information retrieval experiments. In *CTIT Technical Report TR-CTIT-10-15, Centre for Telematics and Information Technology, University of Twente*, pages 1–8.

Myles Hollander and Douglas A. Wolfe. 1973. *Nonparametric Statistical Methods*. John Wiley and Sons, New York.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. Task 3: Cross-level semantic similarity. In *Proceedings of SemEval-2014*.

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of ACL*, pages 1341–1351.

Joaquim Ferreira da Silva, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. 1999. Using localmaxs algorithm for the extraction of contiguous and

[8]Scripts to Index the Anchor ClueWeb12 Dataset are available under request.

Method	Paragraph2Sentence	Sentence2Phrase	Phrase2Word	Word2Sense
Linear (run 3)	0.669	0.671	0.232	0.137
Polynomial (run 1)	0.693	0.665	0.254	0.150
Exponential (run 2)	0.667	0.633	0.180	0.169
Baseline (LCS)	0.527	0.562	0.165	0.109

Table 1: Overall results for the Pearson correlation.

Method	Paragraph2Sentence	Sentence2Phrase	Phrase2Word	Word2Sense
Linear (run 3)	0.688	0.633	0.260	0.124
Polynomial (run 1)	0.666	0.633	0.260	0.126
Exponential (run 2)	0.688	0.633	0.260	0.124
Baseline (LCS)	0.613	0.626	0.162	0.130

Table 2: Overall results for the Spearman's rank correlation.

non-contiguous multiword lexical units. In *Proceedings of EPIA*, pages 113–132.

David Smiley and Eric Pugh. 2009. *Solr 1.4 Enterprise Search Server*. Packt Publishing.

Peter Turney. 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of ECML*, pages 491–502.