Sentence Similarity by Combining Explicit Semantic Analysis and Overlapping N-grams

Hai Hieu Vu[1], Jeanne Villaneau[1], Farida Saïd[2], and Pierre-François Marteau[1]

[1] IRISA, Université de Bretagne Sud (UBS), France
hai-hieu.vu(jeanne.villaneau,pierre-francois.marteau)@univ-ubs.fr,
[2] LMBA, Université de Bretagne Sud, France
farida.said@univ-ubs.fr

Abstract. We propose a similarity measure between sentences which combines a knowledge-based measure, that is a lighter version of ESA (Explicit Semantic Analysis), and a distributional measure, Rouge. We used this hybrid measure with two French domain-orientated corpora collected from the Web and we compared its similarity scores to those of human judges. In both domains, ESA and Rouge perform better when they are mixed than they do individually. Besides, using the whole Wikipedia base in ESA did not prove necessary since the best results were obtained with a low number of well selected concepts.

1 Introduction

Measuring the similarity between sentences is an important task in a variety of applications in natural language processing and related areas, such as information retrieval [3], paraphrase detection [4], text categorization [13] and text summarization [9].

If much work has been presented in the literature for measuring the similarity between long texts and documents, few address the characterization of the similarity between short terms or sentences. P. Achananuparp et al. [1] investigated the efficiency of sentence similarity measures, which are split into three categories by the authors: Word overlap measures, *Tf-Idf* measures and Linguistic Measures as semantic relations and word semantic similarity scores. According to their results, linguistic measures are superior when a low-complexity data-set is considered but, in the presence of "hard" test pairs, for example when one of both sentences can be inferred from the other, most sentence similarity measures do not produce a satisfactory result. More recently, different approaches were tested in the Semantic Textual Similarity task [2] of Semeval 2013, including mixed approaches [5] or UNL (Universal Networking Language) [6].

Measuring similarity requires to define which type of similarity is involved in the corresponding application [17]. In our area of interest, multi-document summarization systems rely widely on similarity measures to both extract the most informative sentences from the original texts and to tackle the issue of semantic redundancy. Since we are interested in extracting the important events of a given domain, the type of similarity between sentences we are seeking for, should focus on the common information and events described in the sentence pairs.

In Section 2, we present a new method for measuring sentence similarity which combines a word co-occurence method and a knowledge-based method. Section 3

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 201–208, 2014.

provides the results of an annotation task involving 7 human participants who scored the similarities in two domain specific datasets composed with 60 selected sentence pairs each. These results are used in Section 4 to evaluate our similarity method. Finally Section 5 summarizes the presented work, draws some conclusions and proposes future perspectives.

2 Proposed System

We propose a similarity measure between sentences which makes use of the knowledge-based model ESA [10,8] together with cosine similarity and the lexical metric Rouge [16]. The similarity score between two sentences is set as a linear combination of Rouge and ESA scores:

$$Score = (1 - p) \times Score_{\text{Rouge}} + p \times Score_{\text{ESA}} \quad (1)$$

where $0 \leq p \leq 1$ is a tuning parameter.

2.1 Lexical Similarity

Rouge (Recall-Orientated Understudy for Gisting Evaluation) is a family of metrics, based on the counts of overlapping units, which was introduced in 2003 [16] to measure the quality of automatically produced summaries through comparison with ideal summaries [15]. We used them in this paper as measures of similarity between any pair of sentences.

We tested four Rouge measures during our experimentations. Rouge1 counts overlapping unigrams, Rouge2 counts overlapping bigrams, Rouge-S enables at most 1 unigram to be skipped inside bigram components and Rouge-SU takes into account both overlapping unigrams and skip-bigrams.

We obtained the best results with Rouge-SU which will be refered to as Rouge in the following.

2.2 Semantic Relatedness

The concept of Explicit Semantic Analysis (ESA) was introduced in 2007 by Gabrilovitch and Markovitch [10,8], to compute the semantic relatedness of words or texts. Since then, it has shown to be highly effective in various applications, among which information retrieval [18], cross-lingual information retrieval [20,21] and text categorization [12,7].

The ESA representation of a word is a vector whose entries reflect its degree of affinity with all documents of a collection, called index collection. A text is represented as the centroid of the vectors representing its words and cosine similarity is used to assess the semantic relatedness between texts. Gabrilovich and Markovitch used Wikipedia articles (concepts) as index documents and a standard *Tf-Idf* weighting scheme for the entries of the vectors.

In [11], Gottron et al. reformulated ESA as a variation of the generalized vector space model (GVSM) [23,22] and showed that ESA essentially captures term correlation

information from the index collection. Larger index collections obviously provide more reliable correlation information. However, stability is reached at some point, so there is no need for excessively large index collections. Furthermore, for applications in a specific domain, there is benefit to take an index collection from the same topic domain while a general topic corpus introduces noise.

In order to build domain-orientated index collections, and prior to ESA, we have retrieved from Wikipedia the K-best concepts that are related to key words of the domain. We use a link mining method, *Pf-Ibf* (Path frequency - Inversed backward link frequency) [19], to score the relatedness of a key word's concept to other concepts in the Wikipedia graph. While *Tf-Idf* analyzes relationships to neighbor articles (only single hop neighbors are considered), *Pf-Ibf* analyzes the relations among nodes in n-hop range. In our experimentations, we considered a 4-hop range neighborhood.

3 Evaluation

We tested our system on two French Web corpora extracted from Wikipedia. The first corpus is about "epidemics" and the second one is about "space conquest".

In the Wikipedia graph, the best related concepts to the key words "space conquest" are mostly named entities which refer to people, countries or names of space vehicles. For the key word "epidemics", we find mostly common nouns, such as names of scientists or diseases.

3.1 Pairs of Sentences

In each corpus, a preliminary human scoring of all pairs of sentences showed a skewed distribution towards 0 (unrelated sentences) and 1 (same topic). In order to construct a set which would reflect a uniform distribution over the range of similarities (0 to 4), we selected a set of sixty sentences as follows: ten sentences, called reference sentences, were selected: they are informative sentences, which contain various important informations of the tested domains. Each of them was associated with six sentences chosen so as to respect the uniform distribution of similarity scores.

A reference sentence (in bold) along with its six associated sentences are presented in Table 1.

3.2 Manual Annotation

We adopted the same annotation procedure as in Li et al. [14]. The participants were asked to rate the similarity between sentences on a scale of 0.0 to 4.0, according to the following definitions:

4: the sentences are completely equivalent;
3: the sentences are mostly equivalent, but some unimportant details differ;
2: the sentences are not equivalent, but they share some parts of information;
1: the sentences are not equivalent, but they are on the same topic;
0: the sentences are unrelated.

Table 1. A reference sentence (in bold) with its associated sentences and their mean similarity scores.

(1) Mars est l'astre le plus étudié du système solaire, puisque 40 missions lui ont été consacrées, qui ont confirmé la suprématie américaine - des épopées Mariner et Viking aux petits robots Spirit et Opportunity (2003 et 2004).
(Mars is the most studied celestial body in the solar system, since 40 missions were dedicated to it, which confirmed the American ascendancy from Marinate and Viking epics to the small robots Spirit and Opportunity (on 2003 and 2004).)
(2) Le 28 novembre 1964, la sonde Mariner 4 est lancée vers Mars, 20 jours après l' échec de Mariner 3.
(On November 28th, 1964, the probe Mariner 4 is launched towards Mars, 20 days after the failure of Mariner 3.)
(3) Les robots Spirit et Opportunity , lancés respectivement le 10 juin 2003 et le 8 juillet 2003 par la NASA , représentent certainement la mission la plus avancée jamais réussie sur Mars.
(Robots Spirit and Opportunity, launched respectively on June 10th, 2003 and July 8th, 2003 by the NASA, represent certainly the most advanced successful mission on Mars.)
(4) Le bilan de l'exploration de Mars est d'ailleurs plutôt mitigé : deux tiers des missions ont échoué et seulement cinq des quinze tentatives d'atterrissage ont réussi (Viking 1 et 2, Mars Pathfinder et les deux MER).
(The assessment of the exploration of Mars is rather mitigated: two thirds of the missions failed and only five out of fifteen landing attempts succeeded (Viking 1 and 2, Mars Pathfinder and both SEA))
(5) Le 6 août 2012, le rover Curiosity a atterri sur Mars avec 80 kg de matériel à son bord.
(On August 6th, 2012, the rover Curiosity landed on Mars with 80 kg of material aboard.)
(6) Arrivé sur Mars en janvier 2004 comme son jumeau Spirit, et prévu comme lui pour fonctionner au moins trois mois, Opportunity (alias MER-B) roule encore et plusieurs de ses instruments répondent présents.
(Arrived on Mars in January 2004 with its twin Spirit, and planned to work at least three months, Opportunity (alias MER-B) still runs and several of its instruments are still operational.)
(7) Mars est mille fois plus lointaine que la Lune et son champ d'attraction plus de deux fois plus intense : la technologie n'existe pas pour envoyer un équipage vers Mars et le ramener sur Terre.
(Mars is a thousand times more distant than the Moon and its gravitation field is more than twice as intense: the technology to send a crew on Mars and get it back on Earth, does not exist .)

pairs	(1)-(2)	(1)-(3)	(1)-(4)	(1)-(5)	(1)-(6)	(1)-(7)
mean score	1.49	2.06	1.86	1.19	1.57	1.1

The participants worked independently and with no time constraint on a web application which was developed to ease the annotation task. For each reference sentence chosen at random, its associated sentences were randomly and successively presented to the annotator. A history of the similarity scores was available and the annotators were free to modify them at any time.

Seven human volunteers, aged 18 to 60, were involved in the annotation task and three of them were experts.

Table 2. Pearson correlation scores between one annotator and the rest of group.

Annotators	1	2	3	4	5	6	7
Correlation (space conquest)	0.872	0.869	0.844	0.941	0.886	0.815	0.855
Standard Deviation (space conquest)	0.586	0.640	0.714	0.364	0.624	0.671	0.568
Correlation (epidemics)	0.862	0.904	0.903	0.931	0.846	0.846	0.806
Standard Deviation (epidemics)	0.544	0.514	0.622	0.367	0.651	0.580	0.617

To investigate the inter-annotator agreement, we compared the scores of each annotator to the averaged scores of the rest of the group. The resulting Pearson correlation scores and standard deviations are given in Table 2. They show that the human raters largely agreed on the definitions used in the scale, even if they found the annotation task quite hard.

4 Results

RESA, that mixes ESA and ROUGE scores, was evaluated using the Pearson correlation coefficient between the system scores and the human scores, as customary in text similarity.

Tables 3 and 4 give Pearson correlations for different values of the tuning parameter p (cf. the formula 1 in Section 2) and for different sizes of the index collection in ESA. The number of reported concepts corresponds to the actual dimension of the representation space, once the ESA inner inverted index has been trimmed. For instance, in Table 3, we selected the 2,000 best related concepts to "space conquest" and we ended up with 1,492 concepts.

Table 3. Pearson correlations between the group of annotators and the system for the "space conquest" dataset. *(p=0:* Rouge *and p=1: ESA)*

$p =$	0	0.10	0.15	0.175	0.20	0.25	0.3	0.35	0.4	0.6	0.8	1
1492 concepts	.800	.814	.819	.821	.8231	.8257	**.8265**	.8251	.821	.777	.682	.554
1857 concepts	.800	.813	.818	.820	.8221	.8247	**.8256**	.8246	.821	.779	.687	.558
3349 concepts	.800	.813	.819	.821	.8225	.8251	**.8259**	.8247	.821	.778	.684	.555

Table 4. Pearson correlations between the group of annotators and the system for the "epidemics" dataset. *($p = 0$:* Rouge *and $p = 1$: ESA)*

$p =$	0	0.10	0.15	0.175	0.20	0.25	0.3	0.35	0.4	0.6	0.8	1
1,016 concepts	.751	.770	.7732	**.7734**	.7726	.768	.760	.749	.735	.666	.591	.525
1,721 concepts	.751	.769	**.7722**	.7720	.7708	.765	.757	.745	.730	.661	.588	.525
3,002 concepts	.751	.770	.7734	**.7735**	.7727	.768	.760	.749	.736	.668	.597	.533
5,094 concepts	.751	.770	.7729	**.7729**	.7720	.767	.759	.748	.734	.666	.595	.531

The main outcome of this experimentation is that, whatever is the size of the index collection, ESA and Rouge perform better when they are combined than they do individually. However, there is a good chemistry to find between them. Rouge always benefits from ESA which in turn, does not benefit from "too much" of Rouge.

With the "space conquest" dataset, the highest correlation score of RESA is 0.8265 and it is achieved for a mixing parameter $p \simeq 0.3$ and about 1,500 concepts. The corresponding scores of the sole Rouge and ESA are 0.800 and 0.554 respectively.

With the "epidemics" dataset, the best performance of RESA is 0.7735 and it is achieved for a mixing parameter $p \simeq 0.175$ and about 3,000 concepts. The corresponding scores of the sole Rouge and ESA are 0.751 and 0.533 respectively.

The best value for the mixing parameter p does not seem to depend on the dimension of ESA representation space but rather seems to depend on the datasets. Indeed, "space conquest" sentences share generally more common words than "epidemics" sentences, which leads in the "epidemics" dataset, to a higher contribution of rouge in the RESA similarity scores.

5 Conclusion

We proposed a similarity measure, (RESA), between sentences which takes advantage of the knowledge-based model ESA and the lexical measure Rouge. The similarity scores provided by RESA are linear combinations of ESA and rouge scores with p and $1 - p$, $0 \leq p \leq 1$ respective weights. When tested on two French datasets collected from the Wikipedia, RESA proved its efficiency and outperforms ESA for p values not exceeding a threshold which seems to depend on the datasets.

Further work is needed to understand what features of the dataset may influence the mixing parameter p (domain, language, lengths of the sentences, their complexity…), and to assess the generalization performance of the RESA measure. Currently, other experiments are conducted on Semeval 2013 data (English language) and in other domains (French language).

The approach presented in this paper is a part of an ongoing work on summarization of domain-oriented French documents. It remains to evaluate the performance of RESA relatively to this specific task.

References

1. Achananuparp, P., Hu, X., Shen, X.: The evaluation of sentence similarity measures. In: Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery. pp. 305–316. DaWaK '08, Springer-Verlag, Berlin, Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-85836-2_29
2. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W.: *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: Semantic Textual Similarity. pp. 32–43. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013), http://www.aclweb.org/anthology/S13-1004

3. Balasubramanian, N., Allan, J., Croft, W.B.: A comparison of sentence retrieval techniques. In: Kraaij, W., de Vries, A.P., Clarke, C.L.A., Fuhr, N., Kando, N. (eds.) SIGIR. pp. 813–814. ACM (2007)
4. Barzilay, R., Elhadad, N.: Sentence alignment for monolingual comparable corpora. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. pp. 25–32. EMNLP '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003), http://dx.doi.org/10.3115/1119355.1119359
5. Buscaldi, D., Le Roux, J., Garcia Flores, J.J., Popescu, A.: 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. pp. 162–168. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013), http://www.aclweb.org/anthology/S13-1023
6. Dan, A., Bhattacharyya, P.: Cfilt-core: Semantic textual similarity using universal networking language. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. pp. 216–220. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013), http://www.aclweb.org/anthology/S13-1031
7. Dasari, D.B., Rao, V.G.: A text categorization on semantic analysis. International Journal of Advanced Computational Engineering and Networking 1(9) (2013)
8. Egozi, O., Markovitch, S., Gabrilovich, E.: Concept-based information retrieval using explicit semantic analysis. ACM Trans. Inf. Syst. 29(2), 8:1–8:34 (Apr 2011), http://doi.acm.org/10.1145/1961209.1961211
9. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. (JAIR) 22, 457–479 (2004)
10. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence. pp. 1606–1611. IJCAI'07, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2007), http://dl.acm.org/citation.cfm?id=1625275.1625535
11. Gottron, T., Anderka, M., Stein, B.: Insights into explicit semantic analysis. In: CIKM'11: Proceedings of 20th ACM Conference on Information and Knowledge Management (2011), http://dl.dropbox.com/u/20411070/Publications/2011-CIKM-Gottron-AS.pdf
12. Gupta, R., Ratinov, L.: Text categorization with knowledge transfer from heterogeneous data sources. In: Proceedings of the 23rd National Conference on Artificial Intelligence – Volume 2. pp. 842–847. AAAI'08, AAAI Press (2008), http://dl.acm.org/citation.cfm?id=1620163.1620203
13. Ko, Y., Park, J., Seo, J.: Automatic text categorization using the importance of sentences. In: In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002. pp. 65–79 (2002)
14. Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K.: Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. on Knowl. and Data Eng. 18(8), 1138–1150 (Aug 2006), http://dx.doi.org/10.1109/TKDE.2006.130
15. Lin, C.: Rouge: a package for automatic evaluation of summaries. pp. 25–26 (2004)
16. Lin, C.Y., Hovy., E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003). Edmonton, Canada (May–June 2003)
17. Lin, D.: An information-theoretic definition of similarity. In: In Proceedings of the 15th International Conference on Machine Learning. pp. 296–304. Morgan Kaufmann (1998)

18. Müller, C., Gurevych, I.: A study on the semantic relatedness of query and document terms in information retrieval. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3. pp. 1338–1347. EMNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009), http://dl.acm.org/citation.cfm?id=1699648.1699680
19. Nakayama, K., Hara, T., Nishio, S.: Wikipedia mining for an association web thesaurus construction. In: In In Proceedings of IEEE International Conference on Web Information Systems Engineering. pp. 322–334 (2007)
20. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: Cross-language plagiarism detection. Lang. Resour. Eval. 45(1), 45–62 (Mar 2011), http://dx.doi.org/10.1007/s10579-009-9114-z
21. Sorg, P., Cimiano, P.: Cross-lingual information retrieval with explicit semantic analysis. In: Working Notes for the CLEF 2008 Workshop (2008), http://www.aifb.kit.edu/images/7/7c/2008_1837_Sorg_Cross-lingual_I_1.pdf
22. Tsatsaronis, G., Panagiotopoulou, V.: A generalized vector space model for text retrieval based on semantic relatedness. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. pp. 70–78. EACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009), http://dl.acm.org/citation.cfm?id=1609179.1609188
23. Wong, S.K.M., Ziarko, W., Wong, P.C.N.: Generalized vector spaces model in information retrieval. In: Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 18–25. SIGIR '85, ACM, New York, NY, USA (1985), http://doi.acm.org/10.1145/253495.253506