October 2019

Parallel sentence alignment from biomedical comparable corpora

Rémi CARDON [a] and Natalia GRABAR [a]
[a] *UMR CNRS 8163 – STL, F-59000 Lille, France*

Abstract. Parallel sentences provide semantically similar information which can vary on a given dimension, such as language or register. Parallel sentences with register variation (like expert and non-expert documents) can be exploited for the automatic text simplification. The aim of automatic text simplification is to better access and understand a given information. In the biomedical field, simplification may permit patients to understand medical and health texts. Yet, there is currently no such available resources. We propose to exploit comparable corpora which are distinguished by their registers (specialized and simplified versions) to detect and align parallel sentences. These corpora are in French and are related to the biomedical area. We treat this task as binary classification (alignment/non-alignment). Our results show that the method we present here can be used to automatically generate a corpus of parallel sentences from our comparable corpus.

Keywords. sentence alignment, text simplification, classification,

1. Introduction

Parallel sentences provide semantically similar information which can vary on a given dimension. The dimension on which the parallelism is positioned can come from many levels, here we are concerned with expert and non-expert register of language. The following pair of sentences illustrates this:

– Expert : *Drugs that inhibit the peristalsis are contraindicated in that situation*
– Non-expert : *In that case, do not take drugs intended for blocking or slowing down the intestinal transit*

Pairs differentiated by their degree of technicality can be used for text simplification. The purpose of text simplification is to provide simplified versions of texts, in order to remove or replace difficult words or information.

Automatic text simplification can be used as a preprocessing step for NLP applications or for producing suitable versions of texts for humans. In this second case, simplified documents are typically created for children [1], or for people with mental or neurodegenerative disorders [2].

Helping patients to better understand medical and health information is an important issue, which motivates our work[3].

In order to perform biomedical text simplification, we propose to collect parallel sentences, which align difficult and simple information. We can exploit an existing monolingual comparable corpus with medical documents in French [4]. The corpus is composed

of documents created for medical professionals and documents created for patients. The purpose of our work is to detect and align parallel sentences from this comparable corpus. We also propose to test what the impact of imbalance on categorization results is: imbalance of categories is indeed a natural characteristic in textual data.

The existing work on searching parallel sentences in monolingual comparable corpora indicates that the main difficulty is that such sentences may show low lexical overlap but be nevertheless parallel. This task is usually explored in general-language corpora and performed as assigning a similarity score from 0 to 5[5]. Among the exploited methods, we can notice lexicon-based methods which rely on similarity of subwords or words from the processed texts or on machine translation [6]; knowledge-based methods which exploit external resources, such as WordNet [7];syntax-based methods which exploit the syntactic modelling of sentences. [8]; or corpus-based methods [9].There is no existing work on building a corpus for text simplification in the biomedical domain.

2. Method

We use the CLEAR comparable medical corpus [4] available online[1] which contains three comparable sub-corpora in French. Documents within these sub-corpora are contrasted by the degree of technicality of the information they contain with typically specialized and simplified versions of a given text. These corpora cover three genres: drug information, summaries of scientific articles, and encyclopedia articles. The *Drugs* corpus contains drug information such as provided to health professionals and patients. This corpus is built from the public drug database[2] of the French Health ministry. The *Scientific* corpus contains summaries of meta-reviews of high evidence health-related articles, such as proposed by the Cochrane collaboration and the simplified versions that they provide[10]. This corpus has been built from the online library of the Cochrane collaboration[3]. The *Encyclopedia* corpus contains encyclopedia articles from Wikipedia[4] and Vikidia[5]. Wikipedia articles are considered as technical texts while Vikidia articles are considered as their simplified versions (they are created for children from 8 to 13 year old). Only articles indexed in the medical portal are exploited in this work.

We exploit a reference dataset with sentences manually aligned by two annotators within this corpus.

2.1. Automatic Detection and Alignment of Parallel Sentences

Automatic detection and alignment of parallel sentences is the main step of our work. The objective is to assess whether two given sentences have the same meaning.

The reference data with aligned sentence pairs, which associate technical and simplified contents, are created manually. We have randomly selected 39 documents from the corpus. The sentence alignment is done by two annotators. This alignment process provides a set of 238 equivalent sentence pairs. The inter-annotator agreement is 0.76

[1] `http://natalia.grabar.free.fr/resources.php#clear`
[2] `http://base-donnees-publique.medicaments.gouv.fr/`
[3] `http://www.cochranelibrary.com/`
[4] `https://fr.wikipedia.org`
[5] `https://fr.vikidia.org`

[11], before consensus. In order to perform the automatic categorization, we also need negative examples, which are obtained by randomly pairing all sentences from all the document pairs except the sentence pairs that are already found to be parallel. Approximately 590,000 non-parallel sentences pairs are created in this way. That high degree of imbalance is the main challenge in our work and we address it in the experimental design (sec. 2.2).

For the automatic alignment of parallel sentences, we use a binary classification model that relies on the random forest algorithm [12]. The implementation we use is the one that is available in scikit-learn [13]. The features we use are the following: number of common non-stopwords; percentage of words from one sentence included in the other sentence, computed in both directions; sentence length ratio; average word length ratio, total number of common character bigrams and trigrams, word-based similarity measure exploits three scores (cosine, Dice and Jaccard) character-based minimal edit distance, word-based minimal edit distance [14], syntactic token-based minimal edit distance; number of common CUI (we generate all the possible subsequences of words in both sentences and check whether they have a corresponding CUI in the UMLS/SNOMED lexicon[16]. We count the number of times that a same CUI appears in both sentences); WAVG and CWASA [17].

2.2. Experimental Design

We work on sentences that are equivalent (they both express the same meaning). 238 equivalent pairs were obtained.

We performed two sets of experiments:

1. We train and test the model with balanced data (we randomly select as many non-aligned pairs as aligned pairs), and then we progressively increase the number of non-aligned pairs until we reach a ratio of 3000:1, which is close to the real data ($\sim$4000:1).
2. Then, for each ratio, we apply the obtained model to the whole dataset and evaluate the results. Note that the training data is included in the whole dataset, we proceed this way because of the low volume of available data.

This means that each model is evaluated twice : once on the test set, and once on the whole dataset. We only report scores for the aligned category. The results that are presented correspond to the mean values over the fifty runs. Finally, we apply the best model on another 30 randomly selected documents and manually evaluate the output.

3. Results

We present the results in Figures 1 and 2: The *x* axis represents the growing of imbalance (the first position is 1 and corresponds to balanced data), while the *y* axis represents the values of Precision, Recall and F-measure. Figure 1 presents the results for the experiments where we evaluate the model on the test set, Figure 2 presents the results for the experiments where we evaluate the model on the whole dataset.

We can observe that the use of balanced data provides very high results, both for Precision and Recall, which are very close to the reference data (> 0.90 performance). These good results in an artifical setting cannot be applied to the real dataset, as is indicated by the starting point in Figure 2

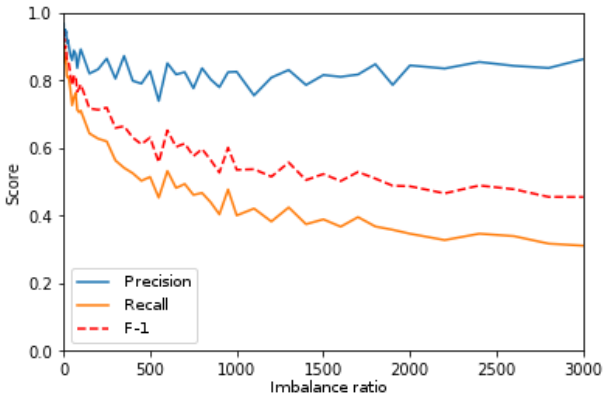

Figure 1. Evaluation on the test set.

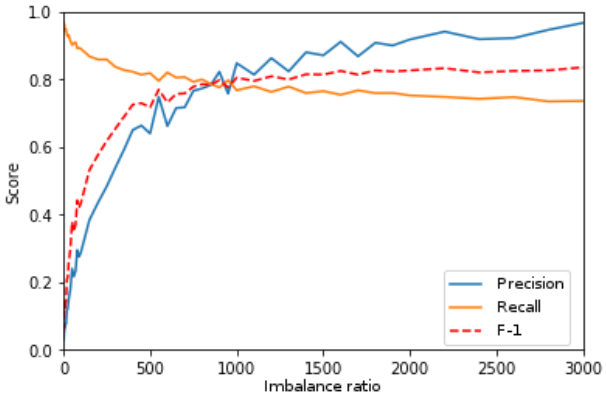

Figure 2. Evaluation on the whole dataset.

4. Discussion

When the model is learned on a substantial degree of imbalance, the Precision score is high when that model is applied to the real data, which has a ratio of about 4,000:1. The recall value is also high, but since two thirds of the aligned sentences have been used for training, that good score should be considered cautiously.

For further evaluation, we randomly selected 30 pairs of documents to evaluate the performances of the models. We used the model that was trained at a ratio of 1200:1. In terms of precision, the model shows 98.75% on all the sentence pairs aligned (80 sentence pairs), including equivalence, inclusions and intersection. Only one result showed two unrelated sentences. Those results show that we have a model that can be used to automatically generate a parallel corpus with reduced noise, from highly imbalanced comparable corpora, for text simplification purposes.

5. Conclusion

We addressed the task of detection and alignment of parallel sentences from a monolingual comparable French corpus. We use the CLEAR corpus, that is related to the biomedical area.

We made observations on the effect of imbalance during training on the performance on the real data. We show that increasing the imbalance during training increases the Precision of the model while still maintaining a stable value for Recall.

We will use that model to generate a corpus of parallel sentences in order to work on the development of methods for biomedical text simplification in French.

References

[1] Tu Thanh Vu, Giang Binh Tran, and Son Bao Pham. Learning to simplify children stories with limited data. In LNCS 8397 Springer, editor, *Intelligent Information and Database Systems*, pages 31–41, 2014.

[2] Ping Chen, John Rochford, David N. Kennedy, Soussan Djamasbi, Peter Fay, and Will Scott. Automatic text simplification for people with intellectual disabilities. In *AIST*, pages 1–9, 2016.

[3] AMA. Health literacy: report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. *JAMA*, 281(6):552–7, 1999.

[4] Natalia Grabar and Rémi Cardon. CLEAR – Simple Corpus for Medical French. In *Workshop on Automatic Text Adaption (ATA)*, pages 1–11, 2018.

[5] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval 2016*, pages 497–511, 2016.

[6] Nitin Madnani, Joel Tetreault, and Martin Chodorow. Re-examining machine translation metrics for paraphrase identification. In *NAACL-HLT*, pages 182–190, 2012.

[7] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Introduction to wordnet: An on-line lexical database. Technical report, WordNet, 1993.

[8] Masashi Tsubaki, Kevin Duh, Masashi Shimbo, and Yuji Matsumoto. Non-linear similarity learning for compositionality. In *AAAI Conference on Artificial Intelligence*, pages 2828–2834, 2016.

[9] Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In *AAAI Conference on Artificial Intelligence*, pages 2786–2792, 2016.

[10] David L. Sackett, William M. C. Rosenberg, Jeffrey A. MuirGray, R. Brian Haynes, and W. Scott Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–2, 1996.

[11] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

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

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics. Doklady*, 707(10), 1966.

[15] Thuppahi Sisira De Silva, Don MacDonald, Grace Paterson, Khokan C. Sikdar, and Bonnie Cochrane. Systematized nomenclature of medicine clinical terms (snomed ct) to represent computed tomography procedures. *Comput. Methods Prog. Biomed.*, 101(3):324–329, March 2011.

[16] Sanja Stajner, Marc Franco-Salvador, Simone Paolo Ponzetto, and Paolo Rosso. Cats: A tool for customised alignment of text simplification corpora. In *Proceedings of the 11th Language Resources and Evaluation Conference, LREC 2018, Miyazaki, Japan, May 7-12*, 2018.