Parallel sentence retrieval from comparable corpora for biomedical text simplification

Rémi Cardon, Natalia Grabar

Parallel sentence retrieval from comparable corpora for biomedical text simplification

Rémi Cardon
UMR CNRS 8163 – STL
F-59000 Lille, France
remi.cardon@univ-lille.fr

Natalia Grabar
UMR CNRS 8163 – STL
F-59000 Lille, France
natalia.grabar@univ-lille.fr

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. Manually created reference data show 0.76 inter-annotator agreement. Our purpose is to state whether a given pair of specialized and simplified sentences is parallel and can be aligned or not. We treat this task as binary classification (alignment/non-alignment). We perform experiments with a controlled ratio of imbalance and on the highly unbalanced real data. Our results show that the method we present here can be used to automatically generate a corpus of parallel sentences from our comparable corpus.

1 Introduction

Parallel sentences provide semantically similar information which can vary on a given dimension. Typically, parallel sentences are collected in two languages and correspond to mutual translations. In the general language, the Europarl (Koehn, 2005) corpus provides such sentences in several pairs of languages. Yet, the dimension on which the parallelism is positioned can come from other levels, such as expert and non-expert register of language. The following pair of sentences (first in expert and second in non-expert languages) illustrates this:

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

Pairs of parallel sentences provide useful information on lexicon used, syntactic structures, stylistic features, etc., as well as the correspondences between the languages or registers. Hence, pairs built from different languages are widely used in machine translation, while pairs differentiated by the register of language can be used for the text simplification. The purpose of text simplification is to provide simplified versions of texts, in order to remove or replace difficult words or information. Simplification can be concerned with different linguistic aspects, such as lexicon, syntax, semantics, pragmatics and even document structure.

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 (Vu et al., 2014), for people with low literacy or foreigners (Paetzold and Specia, 2016), for people with mental or neurodegenerative disorders (Chen et al., 2016), or for laypeople who face specialized documents (Leroy et al., 2013). Our work is related to the creation of simplified medical documents for laypeople, such as patients and their relatives. It has indeed been noticed that medical and health documents contain information that is difficult to understand by patients and their relatives, mainly because of the presence of technical and specialized terms and notions. This situation has a negative effect on the healthcare process (AMA, 1999; Mcgray, 2005; Rudd, 2013). Hence, helping patients to better un-

derstand medical and health information is an important issue, which motivates our work.

In order to perform biomedical text simplification, we propose to collect parallel sentences, which align difficult and simple information, as they provide crucial and necessary indicators for automatic systems for text simplification. Indeed, such pairs of sentences contain cues on transformations which are suitable for the simplification, such as lexical substitutes and syntactic modifications. Yet, this kind of resources is seldom available, especially in languages other than English. As a matter of fact, it is easier to access comparable corpora: they cover the same topics but are differentiated by their registers (documents created for medical professionals and documents created for patients). More precisely, we can exploit an existing monolingual comparable corpus with medical documents in French (Grabar and Cardon, 2018). The purpose of our work is to detect and align parallel sentences from this comparable corpus. We also propose to test what is the impact of imbalance on categorization results: imbalance of categories is indeed the natural characteristics 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. Recently, this task gained in popularity in general-language domain thanks to the semantic text similarity (STS) initiative. Dedicated *SemEval* competitions have been proposed for several years (Agirre et al., 2013, 2015, 2016). The objective, for a given pair of sentences, is to predict whether they are semantically similar and to assign a similarity score going from 0 (independent semantics) to 5 (semantic equivalence). This task is usually explored in general-language corpora (Coster and Kauchak, 2011; Hwang et al., 2015; Kajiwara and Komachi, 2016; Brunato et al., 2016). 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 (Madnani et al., 2012). The features exploited can be: lexical overlap, sentence length, string edition distance, numbers, named entities, the longest common substring (Clough et al., 2002; Zhang and Patrick, 2005; Qiu et al., 2006; Nelken and Shieber, 2006; Zhu et al., 2010);

- knowledge-based methods which exploit external resources, such as WordNet (Miller et al., 1993) or PPDB (Ganitkevitch et al., 2013). The features exploited can be: overlap with external resources, distance between the synsets, intersection of synsets, semantic similarity of resource graphs, presence of synonyms, hyperonyms or antonyms (Mihalcea et al., 2006; Fernando and Stevenson, 2008; Lai and Hockenmaier, 2014);

- syntax-based methods which exploit the syntactic modelling of sentences. The features often exploited are: syntactic categories, syntactic overlap, syntactic dependencies and constituents, predicat-argument relations, edition distance between syntactic trees (Wan et al., 2006; Severyn et al., 2013; Tai et al., 2015; Tsubaki et al., 2016);

- corpus-based methods which exploit distributional methods, latent semantic analysis (LSA), topics modelling, word embeddings, etc. (Barzilay and Elhadad, 2003; Guo and Diab, 2012; Zhao et al., 2014; Kiros et al., 2015; He et al., 2015; Mueller and Thyagarajan, 2016).

There has been work for detection of paraphrases in French comparable biomedical corpora (Deléger and Zweigenbaum, 2009), but there is no work on building a corpus for text simplification in the biomedical domain. Our work is positioned in this area.

In what follows, we first present the linguistic material used, and the methods proposed. We then present and discuss the results obtained, and conclude with directions of future work.

2 Method

We use the CLEAR comparable medical corpus (Grabar and Cardon, 2018) 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

[1] http://natalia.grabar.free.fr/resources.php#clear

and simplified versions of a given text. These corpora cover three genres: drug information, summaries of scientific articles, and encyclopedia articles. We also exploit a reference dataset with sentences manually aligned by two annotators.

2.1 Comparable Corpora

Table 1 indicates the size of the comparable corpus in French: number of documents, number of words (occurrences and lemmas) in specialized and simplified versions. This information is detailed for each sub-corpus: drug information (*Drugs*), summaries of scientific articles (*Scient.*), and encyclopedia articles (*Encyc.*).

The *Drugs* corpus contains drug information such as provided to health professionals and patients. Indeed, two distinct sets of documents exist, each of which contains common and specific information. This corpus is built from the public drug database[2] of the French Health ministry. Specialized versions of documents provide more word occurrences than simplified versions.

The *Scientific* corpus contains summaries of meta-reviews of high evidence health-related articles, such as proposed by the Cochrane collaboration (Sackett et al., 1996). These reviews have been first intended for health professionals but recently the collaborators started to create simplified versions of the reviews (*Plain language summary*) so that they can be read and understood by the whole population. This corpus has been built from the online library of the Cochrane collaboration[3]. Here again, specialized version of summaries is larger than the simplified version, although the difference is not very important.

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). Similarly to the works done in English, we associate Vikidia with Simple Wikipedia[6]. Only articles indexed in the medical portal are exploited in this work. From Table 1, we can see that specialized versions (from Wikipedia) are also longer than simplified versions.

[2] http://base-donnees-publique.medicaments.gouv.fr/

[3] http://www.cochranelibrary.com/

[4] https://fr.wikipedia.org

[5] https://fr.vikidia.org

[6] http://simple.wikipedia.org

Those three corpora have different degrees of parallelism: Wikipedia and Vikidia articles are written independently from each other, drug information documents are related to the same drugs but the types of information presented for experts and laypeople vary, while simplified summaries from the *Scientific* corpus are created starting from the expert summaries.

2.2 Reference Data

The reference data with aligned sentence pairs, which associate technical and simplified contents, are created manually. We have randomly selected 2x14 *Encyclopedia* articles, 2x12 *Drugs* documents, and 2x13 *Scientific* summaries. The sentence alignment is done by two annotators following these guidelines:

1. exclude identical sentences or sentences with only punctuation and stopword difference ;

2. include sentence pairs with morphological variations (e.g. *Ne pas dépasser la posologie recommandée.* and *Ne dépassez pas la posologie recommandée.* – both examples can be translated by *Do not take more than the recommended dose.*);

3. exclude sentence pairs with overlapping semantics, when each sentence brings own content, in addition to the common semantics;

4. include sentence pairs in which one sentence is included in the other, which enables many-to-one matching (e.g. *C'est un organe fait de tissus membraneux et musculaires, d'environ 10 à 15 mm de long, qui pend à la partie moyenne du voile du palais.* and *Elle est constituée d' un tissu membraneux et musculaire. – It is an organ made of membranous and muscular tissues, approximately 10 to 15 mm long, that hangs from the medium part of the soft palate.* and *It is made of a membranous and muscular tissue.*);

5. include sentence pairs with equivalent semantics – other than semantic intersection and inclusion (e.g. *Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation.* and *Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou ralentir le transit intestinal. – Drugs that inhibit peristalsis are contraindicated in*

corpus	*# docs*	*# occ$_{sp}$*	*# occ$_{simpl}$*	*# lemmas$_{sp}$*	*# lemmas$_{simpl}$*
Drugs	11,800x2	52,313,126	33,682,889	43,515	25,725
Scient.	3,815x2	2,840,003	1,515,051	11,558	7,567
Encyc.	575x2	2,293,078	197,672	19,287	3,117

Table 1: Size of the three source corpora. Column headers: number of documents, total number of occurrences (specialized and simple), total number of unique lemmas (specialized and simple)

		Specialized				*Simplified*				*Alignment rate (%)*	
		source		*aligned*		*source*		*aligned*			
	# doc.	*# sent.*	*# occ.*	*# pairs.*	*# occ.*	*# sent.*	*# occ.*	*# pairs.*	*# occ.*	*sp.*	*simp.*
D	12x2	4,416	44,709	502	5,751	2,736	27,820	502	10,398	18	11
S	13x2	553	8,854	112	3,166	263	4,688	112	3,306	20	43
E	14x2	2,494	36,002	49	1,100	238	2,659	49	853	2	21

Table 2: Size of the reference data with consensual alignment of sentences. Column headers: number of documents, sentences and word occurrences for each subset, alignment rate

that situation. and *In that case, do not take drugs intended for blocking or slowing down the intestinal transit.*).

The judgement on semantic closeness may vary according to the annotators. For this reason, the alignments provided by each annotator undergo consensus discussions. This alignment process provides a set of 663 aligned sentence pairs. The inter-annotator agreement is 0.76 (Cohen, 1960). It is computed within the two sets of sentences proposed for alignment by the two annotators.

Table 2 shows the details of the manually aligned set of sentences. Because the three corpora vary in their capacity to provide parallel sentences, we compute their *alignment rate*. The alignment rate for a given corpus is the number of sentences that are part of an aligned pair relative to the total number of sentences. As expected, only a tiny fraction of all possible pairs corresponds to aligned sentences. We can observe that the *Scientific* corpus is the most parallel with the highest alignment rate of sentences, while the two other corpora (*Drugs* and *Encylopedia*) contain proportionally less parallel sentences. Sentences from simplified documents in the *Scientific* and *drugs* corpora are longer than sentences from specialized documents because they often add explanations for technical notions, like in this example: *We considered studies involving bulking agents (a fibre supplement), antispasmodics (smooth muscle relaxants) or antidepressants (drugs used to treat depression that can also change pain perceptions) that used outcome measures including improvement of abdominal pain, global assessment (overall relief of IBS symptoms) or symptom score.* In the *Encylopedia* corpus such notions are replaced by simpler words, or removed. Finally, in all corpora, we observe frequent substitutions by synonyms, like {*nutrition*, *food*}, {*enteral*, *directly in the stomach*}, or {*hypersensitivity*, *allergy*}. Notice that with such substitutions, lexical similarity between sentences is reduced.

The documents are pre-processed. They are segmented into sentences using strong punctuation (*i.e. .?!;:*). We removed, from each subcorpus, the sentences that are found in at least half of the documents of a given corpus. Those sentences are typically legal notices, section titles, and remainders from the conversion of the HTML versions of the documents. The lines that contain no alphabetic characters have also been removed. That reduces the total number of possible pairs for each document pair approximately from 940,000 to 590,000.

2.3 Automatic detection and alignment of parallel sentences

Automatic detection and alignment of parallel sentences is the main step of our work. The unity processed is a pair of sentences. The objective is to categorize the pairs of sentences in one of the two categories:

- alignment: the sentences are parallel and can be aligned;
- non-alignment: the sentences are non-parallel and cannot be aligned.

The reference data provide 663 positive examples (parallel sentence pairs). 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 2.4).

For the automatic alignment of parallel sentences, we first use a binary classification model that relies on the random forests algorithm (Breiman, 2001). The implementation we use is the one that is available in scikit-learn (Pedregosa et al., 2011). Our goal is to propose features that can work on textual data in different languages and registers. We use several features which are mainly lexicon-based and corpus-based, so that they can be easily applied to textual data in other corpora, speacialized areas and languages or transposed on them. The features are computed on word forms (occurrences). The features are the following:

1. *Number of common non-stopwords*. This feature permits to compute the basic lexical overlap between specialized and simplified versions of sentences (Barzilay and Elhadad, 2003). It concentrates on non-lexical content of sentences;

2. *Percentage of words from one sentence included in the other sentence, computed in both directions*. This features represents possible lexical and semantic inclusion relations between the sentences;

3. *Sentence length difference between specialized and simplified sentences*. This feature assumes that simplification may imply stable association with the sentence length;

4. *Average length difference in words between specialized and simplified sentences*. This feature is similar to the previous one but takes into account average difference in sentence length;

5. *Total number of common bigrams and trigrams*. This feature is computed on character ngrams. The assumption is that, at the sub-word level, some sequences of characters may be meaningful for the alignment of sentences if they are shared by them;

6. *Word-based similarity measure exploits three scores (cosine, Dice and Jaccard)*. This feature provides a more sophisticated indication on word overlap between two sentences. Weight assigned to each word is set to 1;

7. *Character-based minimal edit distance* (Levenshtein, 1966). This is a classical acception of edit distance. It takes into account basic edit operations (insertion, deletion and substitution) at the level of characters. The cost of each operation is set to 1;

8. *Word-based minimal edit distance* (Levenshtein, 1966). This feature is computed with words as units within sentence. It takes into account the same three edit operations with the same cost set to 1. This feature permits to compute the cost of lexical transformation of one sentence into another;

9. *WAVG*. This features uses word embeddings. The word vectors of each sentence are averaged, and the similarity score is calculated by comparing the two resulting sentence vectors (Stajner et al., 2018);

10. *CWASA*. This feature is the continuous word alignment-based similarity analysis, as described in (Franco-Salvador et al., 2016).

For the last two features, we trained the embeddings on the CLEAR corpus using word2vec (Mikolov et al., 2013), and the scores are computed using the CATS tool (Stajner et al., 2018).

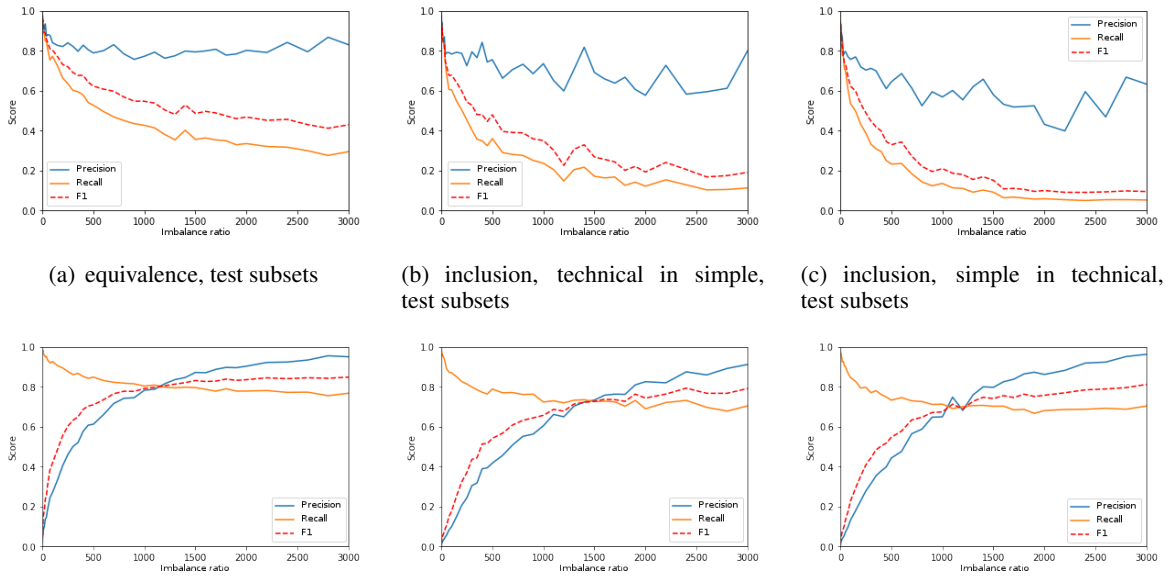
2.4 Experimental design

The set with manually aligned pairs is divided into three subsets:

- *equivalence*: 238 pairs with equivalent semantics,

- *tech in simp*: 237 pairs with inclusion where the content of technical sentence is fully included in simplified sentence, and simplified sentence provides additional content,

- *simp in tech*: 112 pairs with inclusion where the content of simplified sentence is fully included in technical sentence, and technical sentence provides additional content.



(a) equivalence, test subsets

(b) inclusion, technical in simple, test subsets

(c) inclusion, simple in technical, test subsets

(d) equivalence, real data

(e) inclusion, technical in simple, real data

(f) inclusion, simple in technical, real data

Figure 1: Precision, Recall and F-1 for the various experiments and subsets

For each subset, we perform 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 (∼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.

As there is some degree of variability coming with the subset of non-aligned pairs that are randomly selected for the imbalance ratio, every single one of those experiments has been performed fifty times: the results that are presented correspond to the mean values over the fifty runs.

2.5 Evaluation

For evaluating the results, in each experiment we divide the indicated datasets in two parts: two thirds for training and one third for testing. The metrics we use are Recall, Precision and F1 scores. As we are primarily focused on detection of the aligned pairs, we only report scores for that class. Another reason to exclude the negative class and the global score from the observations is that when the data are imbalanced (negative class is growing progressively), misclassifying the positive data has little influence over the global scores, which thus always appear to be high (metrics above 0.99).

Finally, we apply the best model for equivalent pairs on another 30 randomly selected documents and evaluate the output.

3 Presentation and Discussion of Results

We present the results in Figure 1: 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. The results for the three subsets are presented: equivalence (Figures 1(a) and 1(d)), inclusion of technical sentence in simple sentence (Figures 1(b) and 1(e)), and inclusion of simple sentence in technical sentence (Figures 1(c) and 1(f)). Besides, Figures 1(a), 1(b) and 1(c) present the results obtained by training and testing the model on the dataset with the same imbalance ratio (first set of experiments described in section 2.4). As for Figures 1(d), 1(e) and 1(f), they present the results obtained by the models mentioned above that are applied on the whole set of manually annotated data (second set of experiments described in section 2.4).

	equivalence	simp. in tech.	tech. in simp.	intersection	false positives
nb. of pairs	56	10	4	9	1
ratio	70%	12.5%	5%	11.25%	1.25%

Table 3: Breakdown by pair types of the output of the model trained on equivalent pairs with an imbalance ratio of 1200:1 and applied to 30 randomly chosen pairs of documents

The most visible conclusion we can draw from those experiments is that equivalent pairs (Figures 1(a) and 1(d)) are easier to classify than inclusion pairs (the rest of the Figures). Values of both, Precision and Recall, are higher on the equivalence dataset at different imbalance points. For instance, with training on the equally balanced dataset (position 1 on Figure 1(a)), the scores for Precision (0.98) and Recall (0.95) are higher than the scores obtained by the technical in simple dataset (0.96 Precision and 0.94 Recall) and the simple in technical dataset (0.95 Precision and 0.93 Recall) at the same point. For the application to the real data, for ratio 1200:1 – the point where Precision and Recall meet for equivalent pairs, see Figure 1(d) – we obtain 0.81 Precision and 0.81 Recall. At that same ratio, for the technical in simple pairs the scores are 0.65 Precision and 0.73 Recall, and for the simple in technical pairs Precision is 0.73 and Recall is 0.70. This result is positive because the equivalence dataset usually provides the main and the most complete information on transformations required for the simplification. As for the inclusion relations, they cover a large variety of situations which do not necessarily correspond to the searched information. This is illustrated by the unstability of the curves in Figures 1(b) and 1(c), whereas they are smooth in Figure 1(a). The negative examples subset seems to have a quite high influence on the results, which indicates that it is more difficult to draw a clear separation between positive and negative examples. We need to design additional processing steps to be able to classify those pairs in a more efficient way.

We can also observe from Figures 1(a), 1(b) and 1(c) 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). This is true for the three subsets tested (equivalence and inclusions). These good results in an artifical setting cannot be applied to the real dataset, as is indicated by the starting point in Figures 1(d), 1(e) and 1(f). Yet, the imbalance has greater effect on the inclusion datasets, while again the equivalence dataset resists better. An interesting fact is that, 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. This is interesting because it shows that the model is particularly good at discriminating counter-examples. The recall value is also high, but since two thirds of the real data examples have been used for training, that good score should be considered cautiously. We are planning to evaluate the models on a separate set of manually annotated documents. This is still a good result, as during the tests that we performed with other classification algorithms, the models did not successfully recognize the examples they had seen during training.

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 on equivalent pairs. In terms of precision, the model shows 98.75% on all the sentence pairs aligned (80 sentence pairs), including equivalence, inclusions and intersection. Table 3 shows the breakdown of this output in terms of pair types: 70% (56 pairs) are equivalent pairs, 29% (23 pairs) are examples of inclusion (10 simple in technical, 4 technical in simple) and intersection (9), and one pair contains 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 corpus, for text simplification purposes.

4 Conclusion and Future Work

We addressed the task of detection and alignment of parallel sentences from a monolingual comparable French corpus. The comparable aspect is on the technicality axis, as the corpus contrasts technical and simplified versions of information on the same subjects. We use the CLEAR corpus, that is related to the biomedical area.

Several experiments were performed. We divide the task in three subtasks – equivalent pairs,

and inclusion on both directions – and make 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 also find that the task is easier to perform on sentence pairs that have the same meaning, than on sentence pairs where one is included in the other.

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. We will also perform experiments on the general language. Another task we will explore addresses the question on how that model performs with the cross-lingual transfer of descriptors and models.

5 Acknowledgements

This work was funded by the French National Agency for Research (ANR) as part of the *CLEAR* project (*Communication*, *Literacy*, *Education*, *Accessibility*, *Readability*), ANR-17-CE19-0016-01. The authors would like to thank the reviewers for their helpful comments.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo 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 *SemEval 2015*, pages 252–263.

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

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and WeiWei Guo. 2013. *sem 2013 shared task: Semantic textual similarity. In *SEM*, pages 32–43.

AMA. 1999. 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.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *EMNLP*, pages 25–32.

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

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. PaCCSS-IT: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361, Austin, Texas. Association for Computational Linguistics.

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

Paul Clough, Robert Gaizauskas, Scott S.L. Piao, and Yorick Wilks. 2002. METER: Measuring text reuse. In *ACL*, pages 152–159.

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

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Louise Deléger and Pierre Zweigenbaum. 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC)*, pages 2–10, Singapore. Association for Computational Linguistics.

Samuel Fernando and Mark Stevenson. 2008. A semantic similarity approach to paraphrase detection. In *Comp Ling UK*, pages 1–7.

Marc Franco-Salvador, Parth Gupta, Paolo Rosso, and Rafael E. Banchs. 2016. Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *Knowledge-Based Systems*, 111:87 – 99.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *NAACL-HLT*, pages 758–764.

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

Weiwei Guo and Mona Diab. 2012. Modeling sentences in the latent space. In *ACL*, pages 864–872.

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *EMNLP*, pages 1576–1586, Lisbon, Portugal.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard Wikipedia to simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, Denver, Colorado. Association for Computational Linguistics.

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158, Osaka, Japan. The COLING 2016 Organizing Committee.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Neural Information Processing Systems (NIPS)*, pages 3294–3302.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Alice Lai and Julia Hockenmaier. 2014. Illinois-LH: A denotational and distributional approach to semantics. In *Workshop on Semantic Evaluation (SemEval 2014)*, pages 239–334, Dublin, Ireland.

Gondy Leroy, David Kauchak, and Obay Mouradi. 2013. A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *Int J Med Inform*, 82(8):717–730.

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

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

A Mcgray. 2005. Promoting health literacy. *J of Am Med Infor Ass*, 12:152–163.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, pages 1–6.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

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

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

Rani Nelken and Stuart M. Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *EACL*, pages 161–168.

Gustavo H. Paetzold and Lucia Specia. 2016. Benchmarking lexical simplification systems. In *LREC*, pages 3074–3080.

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. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Long Qiu, Min-Yen Kan, and Tat-Seng Chua. 2006. Paraphrase recognition via dissimilarity significance classification. In *Empirical Methods in Natural Language Processing*, pages 18–26, Sydney, Australia.

ED Rudd. 2013. Needed action in health literacy. *J Health Psychol*, 18(8):1004–10.

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

Aliaksei Severyn, Massimo Nicosia, and Alessandro Moschitti. 2013. Learning semantic textual similarity with structural representations. In *Annual Meeting of the Association for Computational Linguistics*, pages 714–718.

Sanja Stajner, Marc Franco-Salvador, Simone Paolo Ponzetto, and Paolo Rosso. 2018. 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*.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Annual Meeting of the Association for Computational Linguistics*, pages 1556–1566, Beijing, China.

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

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

Stephen Wan, Mark Dras, Robert Dale, and Cecile Paris. 2006. Using dependency-based features to take the para-farce out of paraphrase. In *Australasian Language Technology Workshop*, pages 131–138.

Yitao Zhang and Jon Patrick. 2005. Paraphrase identification by text canonicalization. In *Australasian Language Technology Workshop*, pages 160–166.

Jiang Zhao, Tian Tian Zhu, and Man Lan. 2014. ECNU: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. In *Workshop on Semantic Evaluation (SemEval 2014)*, pages 271–277.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *COLING 2010*, pages 1353–1361.