Identification of Parallel Sentences in Comparable Monolingual Corpora from Different Registers

Rémi Cardon, Natalia Grabar

Identification of Parallel Sentences in Comparable Monolingual Corpora from Different Registers

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 aligned sentences provide useful information for different NLP applications. Yet, this kind of data is seldom available, especially for languages other than English. We propose to exploit comparable corpora in French which are distinguished by their registers (specialized and simplified versions) to detect and align parallel sentences. These corpora are related to the biomedical area. Our purpose is to state whether a given pair of specialized and simplified sentences is to be aligned or not. Manually created reference data show 0.76 inter-annotator agreement. We exploit a set of features and several automatic classifiers. The automatic alignment reaches up to 0.93 Precision, Recall and F-measure. In order to better evaluate the method, it is applied to data in English from the *SemEval* STS competitions. The same features and models are applied in monolingual and cross-lingual contexts, in which they show up to 0.90 and 0.73 F-measure, respectively.

1 Introduction

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. Simplification can address needs of people or NLP applications (Brunato et al., 2014). In the first case, simplified documents are typically created for children (Son et al., 1008; De Belder and Moens, 2010; Vu et al., 2014), people with low literacy or foreigners (Paetzold and Specia, 2016), people with mental or neurodegenerative disorders (Chen et al., 2016), or laypeople who face specialized documents (Arya et al., 2011; Leroy et al., 2013). In the second case, the purpose of simplification is to transform documents in order to make them easier to process within other NLP tasks, such as syntactic analysis (Chandrasekar and Srinivas, 1997; Jonnalagadda et al., 2009), semantic annotation (Vickrey and Koller, 2008), summarization (Blake et al., 2007), machine translation (Stymne et al., 2013; Štajner and Popović, 2016), indexing (Wei et al., 2014), or information retrieval and extraction (Beigman Klebanov et al., 2004). Hence, parallel sentences, which align difficult and simple information, provide crucial indicators for the 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. The purpose of our work is to detect and align parallel sentences from comparable monolingual corpora, that are differentiated by their registers. Besides, comparable corpora are easier to obtain. More precisely, we work with texts written for specialists and their simplified versions. We work with corpora in French.

2 Existing Work

In parallel corpora, sentence alignment can rely on empirical information, such as relative length of the sentences in each language (Gale and Church, 1993), or lexical information (Chen, 1993). In comparable corpora, both monolingual and bilingual, sentences present relatively loose common semantics and do not necessarily occur in the same order. It should also be noted that (1) the degree of parallelism can vary from nearly parallel corpora, with a lot of parallel sentences, to *very-non-parallel corpora* (Fung and Cheung, 2004); and that (2) such corpora can contain parallel information at various degrees of granularity, such as documents, sentences or sub-phrastic segments (Hewavitharana and Vogel, 2011). Detection of

parallel sentences in comparable corpora is thus a substantial challenge and requires specific methods.

Several existing works are related to machine translation: bilingual comparable corpora are exploited for creation of parallel and aligned corpora. Usually, these methods rely on three steps:

1. detection of comparable documents using for instance generative models (Zhao and Vogel, 2002) or similarity scores (Utiyama and Isahara, 2003; Fung and Cheung, 2004);

2. detection of candidate sentences, or subphrastic segments, for the alignment using for instance cross-lingual information retrieval (Utiyama and Isahara, 2003; Munteanu and Marcu, 2006), sequence alignment trees (Munteanu and Marcu, 2002), mutual translations (Munteanu and Marcu, 2005; Kumano et al., 2007; Abdul-Rauf and Schwenk, 2009), or dynamic programming (Yang and Li, 2003);

3. filtering and selection of correct extractions using classification (Munteanu and Marcu, 2005; Tillmann and Xu, 2009; Hewavitharana and Vogel, 2011; Ștefănescu et al., 2012), similarity measure of translations (Fung and Cheung, 2004; Hewavitharana and Vogel, 2011), error rate (Abdul-Rauf and Schwenk, 2009), generative models (Zhao and Vogel, 2002; Quirk et al., 2007), or specific rules (Munteanu and Marcu, 2002; Yang and Li, 2003).

In relation with monolingual comparable corpora, the main difficulty is that sentences may show low lexical overlap but be nevertheless parallel. Recently, this task gained in popularity 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 if they are semantically similar and to assign similarity score going from 0 (independent semantics) to 5 (semantic equivalence). This task is usually explored in general-language corpora. 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; Zhao et al., 2014; 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).

These methods and types of features can of course be combined for optimizing the results (Bjerva et al., 2014; Lai and Hockenmaier, 2014; Zhao et al., 2014; Rychalska et al., 2016; Severyn et al., 2013; Kiros et al., 2015; He et al., 2015; Tsubaki et al., 2016; Mueller and Thyagarajan, 2016).

Our objective is close to the second type of works: we want to detect and align parallel sentences from monologual comparable corpora. Yet, there are some differences: (1) we work with corpora related to the biomedical area and not to the general language, (2) we have to state if two sentences have to be aligned (binary statement) and not to compute their similarity score, and (3) we work with data in French which were not exploited for this kind of task yet. To our knowledge, the only work which exploited articles from French encyclopedia performed manual alignment of sentences (Brouwers et al., 2014).

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.

3 Linguistic Material

We use three comparable corpora in French. They are related to the biomedical domain and are contrasted by the technicity of 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 (Sec. 3.1). We also exploit a set of stopwords (Sec. 3.2), and the reference data with sentences manually aligned by two annotators (Sec. 3.3).

3.1 Comparable Corpora

Table 1 indicates the size of the source corpora (number of documents, number of words in specialized and simplified versions). The three corpora are built with French data.

The *Drug* 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[1] of the French Health ministry. These data have been downloaded in June 2017. We can see that the specialized versions of documents provide more word occurrences.

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[2]. The data have been downloaded in November 2017. We can see that specialized version of summaries is also larger than the simplified version, although the difference is not very important.

The *Encyclopedia* corpus contains encyclopedia articles from Wikipedia[3] and Vikidia[4]. Wikipedia articles are considered as technical texts while Vikidia articles are considered as their simplified versions (they are created for children 8 to 13 year old). Similarly to the works done in English, we associate Vikidia with Simple Wikipedia[5]. Only articles related to the medical portal are exploited in this work. These *encyclopedia* articles have been downloaded in August and September 2017. From Table 1, we can see that specialized versions (from Wikipedia) are also longer than simplified versions.

These three corpora are more or less parallel: 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 a lot, while simplified summaries from the *scientific* corpus are created starting from the expert summaries.

3.2 Stopwords

We use a set of 83 stopwords in French, which are mostly grammatical words, like prepositions (*de, et, à, ou (of, and, in, or)*), auxiliary verbs (*est, a (is, has)*) or adverbs (*tout, plusieurs (all, several)*).

3.3 Reference Data

In this section we describe the data that are used for training and evaluation of the automatic sentence alignments.

The reference data are created manually. We have randomly selected 2*14 *encyclopedia* articles, 2*12 *drug* documents, and 2*13 *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 information, in addition to the common semantics;

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

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

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

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

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

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

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

		Specialized				*Simplified*				*Alignment rate (%)*	
		source		*aligned*		*source*		*aligned*			
corpus	*# doc.*	*# pairs.*	*# occ.*	*# pairs.*	*# occ.*	*# pairs*	*# occ.*	*# pairs.*	*# occ.*	*sp.*	*simp.*
Drugs	12*2	4,416	44,709	502	5,751	2,736	27,820	502	10,398	18	11
Scient.	13*2	553	8,854	112	3,166	263	4,688	112	3,306	20	43
Encyc.	14*2	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. (number of sentence pairs and word occurrences for each subset)

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 the peristalsis are contraindicated in 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 indicates the size of the reference data before (*source* columns) and after (*aligned* columns) the alignment. In the two last columns (*Alignment rate*), we indicate the percentage of sentences aligned in each register and corpus. We can observe that *scientific* corpus is the most parallel with the highest alignment rate of sentences from specialized and simplified documents, while the two other corpora (*drugs* and *encylopedia*) contain proportionnally less parallel sentences. Another interesting observation is that 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 in these pairs: {*nutrition*; *food*}, {*enteral*; *directly in the stomach*}, {*hypersensitivity*; *allergy*}, {*incidence*; *possible complications*}. Notice that with such substitutions, lexical similarity between sentences is reduced.

4 Automatic Alignment of Parallel Sentences

As already indicated, our objective is to detect and align parallel sentences within monologual comparable corpora in French. We already have the information on which documents are comparable. So, the task is really dedicated to the alignment of sentences from specialized and simplified versions of documents. The method is composed of several steps: pre-processing of data (Sec. 4.1), generation of features (Sec. 4.2), automatic alignment of sentences (Sec. 4.3), and evaluation (Sec. 4.4).

4.1 Pre-processing of Data

The documents are first pre-processed: they are POS-tagged with TreeTagger (Schmid, 1994), which permits to obtain their lemmatized versions. Then, the documents are segmented into sentences using strong punctuation (*i.e.* *.?!;:*). The same pre-processing and segmentation have been applied when creating the reference data.

4.2 Feature Generation

Our goal is to propose features that can work on textual data in different languages. We use several features which are mainly lexicon-based and corpus-based, so that they can be easily applied to textual data in other languages or transposed to data in other languages. The features are computed on word forms and on lemmas:

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). This feature exploits external knowledge (set of stopwords), which are nevertheless very common linguistic data;

2. Number of common stopwords. This feature also exploits external knowledge (set of stopwords). It concentrates on non-lexical content of sentences;

3. 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;

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

5. 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;

6. 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;

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

8. Word-based similarity measure with the tf*idf weighting of words (Nelken and Shieber, 2006). This feature is similar to the previous one but it also exploits information on context by incorporating the tf*idf weighting (Salton and Buckley, 1988) of words. For this, sentences are considered as documents and documents as corpora. This feature permits to weigh words in a sentence with respect to their occurrences in other sentences of the document;

9. 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;

10. 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.

4.3 Automatic Alignment of Sentences

The task is to find parallel sentences within the whole set of sentences we described in section

3.3. Hence, we have 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 positive examples (663 parallel sentences), while negative examples are obtained by randomly pairing some of the remaining sentences (800 non-parallel sentences) from the same documents.

We use several linear classifiers with their default parameters if not indicated otherwise: Perceptron (Rosenblatt, 1958), Multilayer Perceptron (MLP) (Rosenblatt, 1961), Linear discriminant analysis (LDA) (Fisher, 1936) with the `LSQR` solver, Quadratic discriminant analysis (QDA) (Cover, 1965), Logistic regression (Berkson, 1944), Stochastic gradient descent (SGD) (Ferguson, 1982) with the `log` loss, Linear SVM (Vapnik and Lerner, 1963). We also tested `hinge` and `modified huber` as loss functions with the SGD, and `Eigen` and `SVD` solvers with the LDA, but the results were either lower or very close to the best parameters and we abandoned the idea to use them.

4.4 Evaluation

The training of the system is performed on two thirds of the sentence pairs, and the test is performed on the remaining third. Several classifiers and several combinations of features are tested. Classical evaluation measures are computed: Precision, Recall, F-measure, Mean Square Errors, and True Positives. Our baseline is the combination of length measures with the common words (features 1, 2, 4 and 5). These features are indeed traditionnally exploited in the existing work.

We also evaluate the system on data in English that were released for STS competitions[6]: we use 750 sentence pairs from *SemEval 2012*, 1,500 sentence pairs from *SemEval 2013*, 3,750 sentence pairs from *SemEval 2014*. Each pair of sentences is associated with the similarity score [0;5]. We apply our system to these data in two ways: (1) the system is trained and tested on the STS dataset, and (2) the system is trained on our dataset in French and tested on the STS dataset in English. We assume indeed that the features used and even the models generated can be transposed to data in other languages. For the experiments with the English data, we use the same evaluation measures (Precision, Recall, F-measure, Mean Square Errors, and True Positives). The set of stopwords in English contains 150 entities.

[6] http://ixa2.si.ehu.es/stswiki/index.php/Main_Page

5 Results and Discussion

Classifier	*R*	*P*	*F1*	*MSE*	*TP*
Perceptron	0.87	0.84	0.84	0.63	142
MLP	0.87	0.87	0.86	0.53	167
LDA	0.90	0.90	0.90	0.40	175
QDA	0.89	0.89	0.89	0.45	197
LogReg	0.93	0.93	0.93	0.30	191
SGD	0.87	0.84	0.84	0.84	210
LinSVM	0.81	0.81	0.81	0.74	166

Table 3: Alignment results obtained with different classifiers on French data, test set, whole featureset without tf*idf similarity scores, and non-lemmatized text.

In Table 3, we present the results obtained on French data using the whole set of features (but without the tf*idf similarity scores) on test set, and non-lemmatized texts. The results are indicated in terms of Recall R, Precision P, F-measure F, Mean Square Errors MSE and True positives TP (out of the 221 positive sentence pairs in the test set). We can see that all the classifiers are competitive with F-measure above 0.80. Overall, several classifiers (LDA, QDA, LogReg, LinSVM) provide stable results, for which we indicate the evaluation scores obtained in one iteration. Other classifiers (Perceptron, MLP, SGD) provide fluctuating results, and we indicate then the average scores obtained after 20 iterations. Another positive observation is that Precision and Recall values are well balanced. Logistic regression seems to be the best classifier for this task, with Precision, Recall and F-measure at 0.93. This classifier is used for the experiments described in the next sections.

We first present and discuss the exploitation of various featuresets on French data (Sec. 5.1), and then the exploitation of the features and models on the STS data in English in monolingual (Sec. 5.2) and cross-lingual (Sec. 5.3) contexts. As our final objective (text simplification in French) and the data we work on (French texts from the biomedical domain) are different from the STS context, we believe it should be noted that there are intrin-

sic limitations as to the comparison we can make.

5.1 Different Featuresets

Feature set	*R*	*P*	*F1*	*MSE*	*TP*
BL	0.87	0.87	0.86	0.54	173
S	0.84	0.84	0.84	0.64	174
L	0.79	0.78	0.78	0.86	146
N	0.89	0.88	0.88	0.48	168
L+S	0.88	0.88	0.88	0.48	170
L+N	0.91	0.91	0.91	0.37	187
S+N	0.91	0.91	0.91	0.37	183
BL+L	0.90	0.90	0.90	0.40	184
BL+S	0.89	0.89	0.89	0.46	180
BL+N	0.91	0.91	0.91	0.35	187
BL+L+S	0.90	0.90	0.90	0.40	184
BL+L+N	0.93	0.93	0.93	0.29	191
BL+S+N	0.91	0.91	0.91	0.36	189
L+S+N	0.91	0.91	0.91	0.36	189
BL+L+S+N	0.93	0.93	0.93	0.29	191

Table 4: Alignment results obtained with various featuresets, logistic regression, non-lemmatized text.

The purpose of these experiments is to detect the most suitable combinations of features. We present the results obtained on our data. We distinguish four sets of features, which are used in isolation and in various combinations. We indicate the corresponding numbers from section 4.2 between brackets :

1. BL: baseline (1, 2, 3, 4 5);
2. L: Levenshtein-based features (9, 10);
3. S: similarity-based features (7, 8);
4. N: ngram-based features (6).

Contrary to the previous work (Nelken and Shieber, 2006; Zhu et al., 2010), the tf*idf weighting of words is not efficient on our data. For this reason, this set of features was not used in the experiments.

The results are presented in Table 4. The lowest results are obtained with the Levenshtein-based features (F-measure 0.78), they are followed by the similarity-based features (F-measure 0.84). We obtain 0.86 F-measure with the baseline. Other combinations indicate that each set of features exploited is useful to gain efficiency for this task. Hence, the best results are obtained with the combination BL+L+N and with the whole set of features (BL+L+S+N), which shows 0.93 F-measure. We use the whole set of features for the experiments with the STS dataset.

5.2 Classification of the STS Sentence Pairs

STSset score	*R*	*P*	*F1*	*MSE*	*TP*
STS2012 2.5	0.82	0.82	0.82	0.71	477
STS2012 3.5	0.74	0.74	0.74	1.04	277
STS2012 4.5	0.79	0.81	0.78	0.74	37
STS2013 2.5	0.73	0.73	0.73	1.09	176
STS2013 3.5	0.78	0.78	0.78	0.87	96
STS2013 4.5	0.89	0.93	0.90	0.29	2
STS2014 2.5	0.75	0.76	0.75	0.97	653
STS2014 3.5	0.70	0.71	0.71	1.17	306
STS2014 4.5	0.89	0.93	0.90	0.29	2

Table 5: Alignment results obtained on the STS data in English, test set, whole featureset, logistic regression, non-lemmatized text and training on the STS data.

In this set of experiments, the classification model is trained and tested on the STS reference data in English. Our assumption is that the features exploited are transferable from one language to another. The reference data and categories in English and in French differ. One difference is that the STS pairs of sentences are scored from 0 to 5 according to their similarity, while in French we do binary classification (a given pair of sentences should be aligned or not). To make the two datasets comparable, we propose to transform the STS scoring in binary categories. We test similarity thresholds within the interval [2.5;4.5] by step of 0.5, which permits not to consider identical sentences (scores close to 5) and very distant sentences (scores lower than 2.5). As indicated in Table 5, we obtain up to 0.90 F-measure with the similarity threshold 4.5 on data from 2013 and 2014, while in 2012 the best F-measure (0.82) is obtained with the similarity score 2.5. It is difficult to compare our results with those of the participating teams and already published results because our categories and evaluation differ from the STS protocols – we rate sentence pairs as either aligned or not aligned, while STS offers a scale from 0 to 5. Yet, the MSE rate (0.308) published by one of the top participants in 2014 (Bjerva et al., 2014) indicates that our MSE rate is improved, as it is at 0.29 on the 2014 data.

5.3 Cross-lingual Classification of the STS Sentence Pairs

STSset score	*R*	*P*	*F1*	*MSE*	*TP*
STS2012 2.5	0.83	0.81	0.82	0.19	1378
STS2012 3.5	0.74	0.72	0.71	0.28	1035
STS2012 4.5	0.81	0.49	0.52	0.51	413
STS2013 2.5	0.74	0.74	0.74	0.26	523
STS2013 3.5	0.78	0.73	0.74	0.27	396
STS2013 4.5	0.92	0.57	0.67	0.43	88
STS2014 2.5	0.74	0.72	0.73	0.28	1688
STS2014 3.5	0.72	0.69	0.69	0.31	1216
STS2014 4.5	0.88	0.54	0.61	0.46	384

Table 6: Alignment results obtained on the STS data in English, test set, whole featureset, Logistic regression, non-lemmatized text and training on the French data.

In this set of experiments, the classification model is trained on French data and tested on the STS data in English. Here, our assumption is that the models generated on one language can be transferable to another language in order to detect parallel sentences. Here as well, we test several similarity thresholds. As we can see in Table 6, in this cross-lingual experiment, the best F-measures are obtained with the score 2.5 in 2012 (0.82) and in 2014 (0.73), and with scores 2.5 and 3.0 in 2013 (0.74). These thresholds indicate that the models generated on our French data can be exploited on the STS data in English quite efficiently and that the features that are used show cross-lingual relevance for the French-English language pair. These results also indicate that, for the targeted task of text simplification, we need quite a strong similarity between sentences.

6 Conclusion and Future Work

In this work, we proposed to address the task of detection and alignment of parallel sentences from monolingual comparable corpora in French. The comparable dimension is due to the technicality of documents, which contrast specialized and simplified versions of documents and sentences. We use three corpora which are related to the biomedical area. Several features and classifiers and exploited. Our results reach up to 0.93 F-measure on the French data, with a very good balance between Precision and Recall. Linear regression appears to be the best classifier for this task. Our approach is then tested on the STS data in English, such as proposed by several *SemEval* competitions between 2012 and 2014. We first test the features, with training and testing done on the STS data. This gives up to 0.90 F-measure with the 4.5 similarity threshold. Then, we test the models: they are generated on the French data and tested on the STS data. This gives 0.82 F-measure. We assume that the proposed approach (features and classifiers) show a good transferability to another language. This is a good point because it validates our approach on data from another language.

In future, we plan to exploit the best models generated in French for enriching the set of parallel sentences. This will permit to prepare data necessary for the developement of simplification methods for French. Parallel sentences may also be helpful for othe NLP applications. Other directions for future work are concerned with the exploitation of other features for the alignment of sentences, such as use of word embeddings to smooth lexical variation or exploitation of external knowledge. Besides, our appoach will be further evaluated on data from other languages.

7 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

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve smt performance. In *European Chapter of the ACL*, pages 16–23.

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.

Diana J. Arya, Elfrieda H. Hiebert, and P. David Pearson. 2011. The effects of syntactic and lexical complexity on the comprehension of elementary science texts. *Int Electronic Journal of Elementary Education*, 4(1):107–125.

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

Beata Beigman Klebanov, Kevin Knight, and Daniel Marcu. 2004. Text simplification for information-seeking applications. In Robert Meersman and Zahir Tari, editors, *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*. Springer, LNCS vol 3290, Berlin, Heidelberg.

Joseph Berkson. 1944. Application of the logistic function to bio-essay. *Journal of the American Statistical Association*, 39:357–365.

Johannes Bjerva, Johan Bos, Rob van der Goot, and Malvina Nissim. 2014. The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. In *Workshop on Semantic Evaluation (SemEval 2014)*, pages 642–646, Dublin, Ireland.

Catherine Blake, Julia Kampov, Andreas Orphanides, David West, and Cory Lown. 2007. Query expansion, lexical simplification, and sentence selection strategies for multi-document summarization. In *DUC*.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas Francois. 2014. Syntactic sentence simplification for French. In *PITR workshop*, pages 47–56.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2014. Defining an annotation scheme with a view to automatic text simplification. In *CLICIT*, pages 87–92.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge Based Systems*, 10(3):183–190.

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.

Stanley F Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Annual Meeting of the Association for Computational Linguistics*, pages 9–16.

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.

Thomas M. Cover. 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14(3):326–334.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Workshop on Accessible Search Systems of SIGIR*, pages 1–8.

Thomas S. Ferguson. 1982. An inconsistent maximum likelihood estimate. *Journal of the American Statistical Association*, 77(380):831–834.

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

Ronald A. Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Pascale Fung and Percy Cheung. 2004. Mining very non-parallel corpora: Parallel sentence and lexicon extraction vie bootstrapping and em. In *Conference on Empirical Methods in Natural Language Processing*, pages 57–63.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Comp Linguistics*, 19(1):75–102.

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

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.

Sanjika Hewavitharana and Stephan Vogel. 2011. Extracting parallel phrases from comparable data. In *4th Workshop on Building and Using Comparable Corpora*, pages 61–68.

Siddhartha Jonnalagadda, Luis Tari, Jörg Hakenberg, Chitta Baral, and Graciela Gonzalez. 2009. Towards effective sentence simplification for automatic processing of biomedical text. In *NAACL HLT 2009*, pages 177–180.

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.

Tadashi Kumano, Hideki Tanaka, and Takenobu Tokunaga. 2007. Extracting phrasal alignments from comparable corpora by using joint probability SMT model. In *Int Conf on Theoretical and Methodological Issues in Machine Translation*.

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.

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

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.

Dragos Stefan Munteanu and Daniel Marcu. 2002. Processing comparable corpora with bilingual suffix trees. In *EMNLP*, pages 289–295.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *COLING-ACL*, pages 81–88.

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.

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.

Chris Quirk, Raghavendra U. Udupa, and Arul Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *Machine Translation Summit XI*, pages 377–384.

Frank Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

Frank Rosenblatt. 1961. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington DC.

Barbara Rychalska, Katarzyna Pakulska, Krystyna Chodorowska, WojciechWalczak, and Piotr Andruszkiewicz. 2016. Samsung Poland NLP team at SemEval-2016 task 1: Necessity for diversity; combining recursive autoencoders, wordnet and ensemble methods to measure semantic similarity. In *SemEval-2016*, pages 614–620.

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.

Gerard Salton and Chris Buckley. 1988. Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Int Conf on New Methods in Language Processing*, pages 44–49.

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.

Ji Y. Son, Linda B. Smith, and Robert L. Goldstone. 1008. Simplicity and generalization: Short-cutting abstraction in children's object categorizations. *Cognition*, 108:626–638.

Sara Stymne, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 2013. Statistical machine translation with readability constraints. In *NODALIDA*, pages 1–12.

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.

Christoph Tillmann and Jian-Ming Xu. 2009. A simple sentence-level extraction algorithm for comparable data. In *Companion Vol. of NAACL HLT*.

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.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *Annual Meeting of the Association for Computational Linguistics*, pages 72–79.

Vladimir Vapnik and A. Lerner. 1963. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:709–715.

David Vickrey and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *Annual Meeting of the Association for Computational Linguistics-HLT*, pages 344–352.

Sanja Štajner and Maja Popović. 2016. Can text simplification help machine translation? *Baltic J. Modern Computing*, 4(2):230–242.

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.

Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. 2014. Simconcept: A hybrid approach for simplifying composite named entities in biomedicine. In *BCB '14*, pages 138–146.

Christopher C. Yang and Kar Wing Li. 2003. Automatic construction of English/Chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, 54(8):730–742.

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

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *IEEE Int Conf on Data Mining*, pages 745–748.

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)*, page 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.

Dan Ştefănescu, Radu Ion, and Sabine Hunsicker. 2012. Hybrid parallel sentence mining from comparable corpora. In *16th EAMT Conference*, pages 137–144.