Building and Exploiting Domain-Specific Comparable Corpora for Statistical Machine Translation

Rahma Sellami, Fatiha Sadat and Lamia Hadrich Beluith

Abstract In this paper we address the problem of mining domain-specific comparable and parallel data to improve the accuracy of a Statistical Machine Translation system. First, we present a novel strategy for building domain-specific comparable corpora from Wikipedia. Our strategy exploits the categorization and the multilingualism of Wikipedia documents in order to extract domain-specific comparable corpora. Second, we describe a combined anchor-point-based method for comparable sentences alignment. Third, we present a compositional-based approach for parallel phrase mining. We conducted multiple evaluations to qualify the extracted comparable and parallel data. Applied to Arabic and French languages pair, we extract 81 domain-specific comparable and parallel corpora. The extracted parallel data are used to adapt an Arabic to French domain-generic SMT system to a specific domain one. This additional training data provided significant improvements of the translation quality in terms of BLEU and OOV scores over the baseline system.

Keywords Comparable corpora • Compositional-based approach
Domain-specific • SMT • Anchor point

1 Introduction

Statistical Machine Translation (SMT) use comparable or parallel corpora as essential resources to train translation models. These corpora are widely available for general-domain but not for specific domains such as art, society and media.

R. Sellami (✉) · L.H. Beluith
ANLP Research Group, MIRACL Laboratory, Sfax University, Sfax, Tunisia
e-mail: rahma.sellami@fsegs.rnu.tn

L.H. Beluith
e-mail: l.belguith@fsegs.rnu.tn

F. Sadat
University of Quebec in Montreal, Montreal, Canada
e-mail: sadat.fatiha@uqam.ca

K. Shaalan et al. (eds.), *Intelligent Natural Language Processing: Trends and Applications*, Studies in Computational Intelligence 740,
https://doi.org/10.1007/978-3-319-67056-0_31

Systems specialized in specific domains require in-domain training data to give the best performance.

Very productive methods for creating domain-generic comparable and parallel corpora have been proposed [1, 17, 24]. Nevertheless, very few researches have been done for domain-specific data [2]. In this paper, we first present a novel strategy, based on category tags and inter-language links, for mining many domain-specific comparable corpora from Wikipedia. Then, a combined anchor point-based-method is proposed for comparable sentences mining. Anchor points are elements aligned with trust and which methods can be based to reduce the search space in order to align their neighbor [4]. Various types of anchor points are proposed and combined for comparable sentences alignment and thus complete a compositional-based approach for parallel phrase mining. The compositional translation based approach consists of the fact that the translation of an expression is a function of the translation of the parts [11]. This approach has proved its effectiveness for bilingual lexicon mining from comparable corpora [7]. Nevertheless, no works have been done in the field of parallel phrase mining. In this paper, we propose to exploit the compositional translation based approach for parallel phrase mining from comparable corpora.

This paper is organized as follows. Section 2 presents previous works on mining domain-specific comparable and parallel corpora. Section 3 describes the proposed strategy for domain-specific comparable corpora extraction from Wikipedia. Section 4 illustrates an anchor-point-based method for comparable sentences alignment. Section 5 presents a compositional-based approach for parallel phrase mining. Section 6 evaluates the resulting comparable and parallel domain-specific corpora applied to Arabic and French languages pair. The last section concludes the present paper with future extensions.

2 Related Works

Very few works have studied the mining of domain-specific parallel corpora from domain-generic multilingual resources. Most of these works are based on information retrieval approaches. Plamada and Volk [22] proposed an approach for mining Alpine domain parallel corpora from Wikipedia. They exploited inter-language links to extract comparable domain-generic articles. The extracted corpus is subsequently used for information retrieval queries aiming to identify the articles belonging to the Alpine domain. Parallel sentences are then selected by means of similarity metric [34] developed a configurable Focused Monolingual Crawler for collecting domain-specific corpora from the Web.Then, they presented a method for extracting bilingual named entities, phrases and sentences from the collected corpora. Pal et al. [20] designed a crawler to collect comparable corpora from Wikipedia, based on an initial seed keyword list and inter-language links. Textual entailment techniques are then used to extract parallel phrases from these comparable corpora. The parallel text fragments extracted thus were able to bring some improvements in the performance of an existing MT system on the tourism domain. Gamallo and Loopez [10] proposed a strategy to extract CorpusPedia, bilingual comparable corpora, from Wikipedia. They specified some categories to make the collected corpus comparable according to some specific topics. Also, a measure of comparability is used to verify whether the corpora are lowly or highly comparable. The difference with respect to our strategy is that they only consider the articles associated to one specific category and not to an entire domain.

Barrón-Cedeño et al. [2] proposed a simple model for extracting comparable corpora from Wikipedia based on the category graph. Our strategy for domain-specific comparable corpora mining is close to [2]. The difference between our proposal and the Barrón-Cedeño et al.'s proposal relies in the fact that we explore the whole category graph. However, [2] used a stopping criterion based on a domain vocabulary list. We assume that the vocabulary list could not be complete and this hypothesis can reduce the coverage of the resulting comparable corpus. Recently, [6] proposed an integrated system to extract both parallel sentences and fragments from comparable corpora. They first applied parallel sentence extraction to identify parallel sentences from comparable sentences. Then they extracted parallel fragments from the comparable sentences. Parallel sentence extraction is based on a filter and a classifier. Chu et al. [6] improved this method by proposing a novel filtering strategy and three novel feature sets for classification. They demonstrated that the extracted parallel data significantly improves SMT performance. Wolk et al. [38] proposed a method of automatic web crawling in order to build topic-aligned comparable corpora. They developed methods of obtaining parallel sentences from comparable data and proposed methods of filtration of corpora capable of selecting inconsistent or only partially equivalent translations. Evaluation of the quality of the created corpora was performed by analyzing the impact of their use on statistical machine translation systems.

3 Domain-Specific Comparable Corpora Building

The domain-specific comparable corpora strategy we propose in this paper is designed to exploit Wikipedia categorization and inter-language links.

Wikipedia articles form a network of semantically related terms called Wikipedia Articles Graph (WAG), while the categories are organized in a taxonomy-like structure called Wikipedia Category Graph (WCG) [39]. Articles are usually not placed in the most general category they logically belong to and are tagged as a sub-category thereof which forms a Category-Article Graph (CAG). This is the concatenation of WAC and WCG (Fig. 1). Cycles and shortcuts occur among the different categories. We first extracted the main categories of Wikipedia, the sub-categories and all articles associated to each category. We extracted 6 main categories (art, society, science, technology, space and time, people) and a total

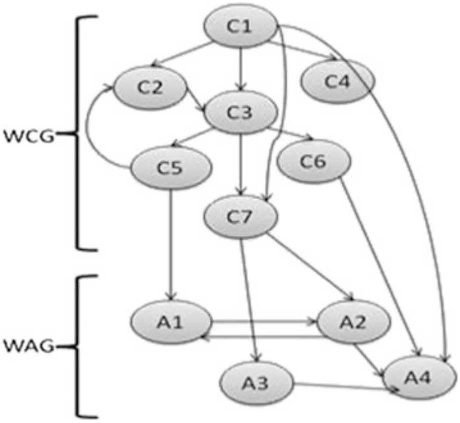


Fig. 1 Category-Article Graph (CAG)

of 81 sub-categories (architecture, film, language, media, tourism, biology, agriculture, robotics, etc.). Thus, we constructed a CAG for each sub-category. The root is the domain name and the endpoints are the titles of the articles associated to one domain. Once the CAG is constructed for each domain, we parsed the graph and the Arabic and French Wikipedia dumps and we extract bilingual articles (related by inter-language links). The output is a set of comparable corpora classified in Wikipedia domains and aligned at article level.

4 Combined Anchor-Point-Based Method for Comparable Sentences Alignment

The main idea of the comparable sentences alignment process is to find correlative elements, also called anchor points, in comparable sentences.

Anchor points can be structural information associated with the document title, subtitle, caption, etc. [25]. They can also be lexical [14] and be extracted based on a bilingual dictionary [12] or transliterations properties of languages. Prochasson et al. [23] have defined some properties of the anchor points: they should be easily identified, relevant regarding corpora topics and not polysemous.

We start with some pre-processings. It consists of tokenization, normalization, lemmatization of source and target sides, truecasing the French letters and stop-words removing.

We propose to combine four types of anchor points for comparable sentence alignment.

- **Word frequency**

Word frequencies have been used in many previous works in information retrieval. Lardilleux and Lepage [15] investigated the use of hapaxs (words that occur only once in a single document) for word alignment and concluded that they can safely be aligned in most cases. This notion is also used in [22] for parallel document alignment. In contrast, [9] exploit high frequency words and their translations for aligning noisy parallel corpora.

In this paper, we propose to exploit words occurring less than four times and the most frequent words as anchor points for comparable sentences alignment.

- **Bilingual fragments**

We extract article titles related by inter-language links. Also, files, images and videos in Wikipedia are often stored in a central source across different languages. This allows the identification of captions, which are most of the time parallel [29, 35].

We exploit the fact that these titles and captions will appear in the text body of articles. If a pair of candidate sentences contains such bilingual fragments (title or caption), it is most likely comparable.

- **Named entities**

Named Entities (NEs) are expressions commonly used and are frequent in all kinds of texts. Bilingual NEs were previously used in many works. Samy et al. [27] used bilingual NE as anchor points for parallel corpus alignment. Semmar and Saadane [32] exploited NE transliteration to improve the results of a linguistic word alignment approach from parallel text corpora.

We make the following assumption: if a NE co-occurs in two sentences, they are very likely to talk about the same event. In this work, two sources of NE translations are exploited. Wikipedia and United Nation corpora are used for person, location and organization named entity translation mining [28, 31].

- **Cognates**

Cognates, words that have similar spelling between two languages, are easy to discover in similar spelling languages. Otherwise, authors use transliteration to close language spelling. Many authors used cognate-based features for alignment of parallel or comparable corpus [1, 32].

We propose two methods for cognate's detection from bilingual sentences. The first one selects cognate's type foreign language words, digits, alphanumerical symbols or punctuation marks. These strings appear reliably in comparable sentences. The second method is based on the transliteration of Arabic words; it can select only words of similar length with a large number of common characters regardless of the order. For this purpose, we define two scores Distance_Score and Length_Score:

$$\text{Distance\_Score} = \frac{\text{editDistance}(\text{ar}, \text{fr})}{(\max(|\text{ar}|, |\text{fr}|))} \tag{1}$$

$$\text{Length\_Score} = \frac{\max(|\text{ar}|, |\text{fr}|)}{\min(|\text{ar}|, |\text{fr}|)} \tag{2}$$

where max(|ar|, |fr|) is the number of characters of the longest string and Min(|ar|, |fr|) is the number of characters of the smallest string. EditDistance is the Editex technique [40], based on a variant of Levenshtein edit distance algorithm [16]. Editex combines edit distance with the use of a group of similar letters (aeiouy, bp, ckq, dt, lr, mn, gj, fpv, sxz, csz); such letters in a similar group frequently correspond to a similar pronunciation. As in Levenshtein distance, the minimal number of insertions, deletions, and replacements necessary to transform one string to another is computed. However, edits that replace a letter with another letter from a different group are weighted more heavily, and deletions of letters that are frequently silent (h and w) are weighted less heavily than other deletions. According to these scores, two words are cognates if Distance_Score is lower than 0.6 and Length_Score is lower than 1.5. These two values are fixed empirically.

All cited anchor points are combined for comparable sentences alignment. At the end of this step, a similarity score Sim_Anch is attributed for each pair of sentences.

$$\text{Sim\_Anch} = \sum_{i=1}^{n=4} \text{Count}(\text{Anchor}(i)) \tag{3}$$

Only pairs of sentences with a similarity score equal or greater than a threshold Anch are included in the sentence aligned comparable corpus.

5 Compositional-Based Approach for Parallel Fragment Generation

In this section, a new approach for mining parallel phrases from comparable sentences pairs based on the compositional translation [11] is proposed.

The input of this step is pairs of pre-processed comparable sentences (in form of bag of lemma of lexical words). Phrase generation, phrase translation, re-composition and filter steps are executed in order to generate fragment translations.

- **Phrase generation**

In order to generate source and target phrases from comparable sentences we extract all n-grams up to length 5 from each lemmatized sentence.

- **Phrase translation**

Compositional translation consists of translating each lemma in the source phrase. The translation step considers all alternative translations based on lexicons and anthologies. For each lemma in the source phrase, the following steps are considered.

1. First, semantic relations (such as synonyms and hyponyms) are defined using an Arabic ontology [8].
2. Second, lemma and its semantic relations are translated based on bilingual lexicons. For this purpose, we use GLOB-LEX, a bilingual lexicon based on many resources: a bilingual lexicon extracted from Wikipedia titles and based on a hybrid approach [30], anchor points which contribute to the comparable sentences mining and some dictionaries (the universal dictionary, wiktionary and Omegawiki). All these data are in dictionary format (lemma). GLOB-LEX is composed of 2 219 509 pairs of terms. We consider all hypothesis translations.
3. Third, semantic relations for each translation are added based on a French ontology [3].
4. Finally, all generated translations are revised to delete any duplicated translations.

- **Re-composition**

We re-compose the translation candidates of a fragment, taking into account all possible combinations. In order to overcome distortion phenomena, fragments are treated as bags of translated lemmas. In the end of this step, N translation candidates are produced.

$$\mathrm{N} = \prod_{i=1}^{i=m} T_i \tag{4}$$

where Ti is the number of generated translations of a source lemma i. m is the number of lemmas in the source fragment. Ti = 1 if no translation was generated for a lemma i.

- **Translation filter**

This step consists of matching the sequence of translated lemmas with tokens lemmas in the comparable sentence (in target language). Ideally, we should select only pairs of sequences that co-occur exactly in the comparable sentences. But, due to translation phenomenon (insertion and deletion) that appear in comparable sentences and the fact that lexicons cannot cover all lemma in our comparable corpus, we propose to accept pairs of sequences containing some insertions and deletions.

The process of translation filter consists of the following steps. Given a sequence of lemmas in the source language and various translation candidates produced by the previous step, we select the best translation sequence based on the target side of the comparable sentence and using a lemma-overlap score. This score is calculated based on the translated lemmas and all sequences of lemmas generated from the target comparable sentence.

6 Evaluation

6.1 Domain-Specific Comparable Corpora Evaluation

In this section, we conduct an evaluation of the degree of comparability of comparable corpora based on a quantitative comparability measure C_{LG}.

Note that C_{LG} is a comparability measure, proposed by [18] based on vocabulary similarity. Table 1 shows the degree of comparability of Arabic–French domain-specific documents extracted from Wikipedia. First, it presents the number of Arabic documents, French documents and bilingual comparable documents. Second, it presents the percentage of comparable documents in many cases: C_{LG} is equal to 0, between 0 and 0.2, between 0.2 and 0.3 and greater than 0.3. For example, considering the Eating topic, we notice that 73.68% of documents have a comparability measure, C_{LG}, equal to 0 (these documents can be characterized as semi-comparables). Whereas, C_{LG} is between 0 and 0.2 in 13.53% of documents and it is between 0.2 and 0.3 in 9.77% of documents. While, only 3% of documents have a comparability measure which is greater than 0.3. These documents can

Table 1 Comparability evaluation of domain-specific comparable corpora

Domain	# ar doc	# fr doc	# comp doc	$C_{LG} = 0$ (%)	$C_{LG} > 0$ $C_{LG} < 0.2$ (%)	C_{LG} >=0.2 $C_{LG} < 0.3$ (%)	C_{LG} >=0.3 (%)
Eating	235	145	133	73.68	13.53	9.77	3.01
Media	893	854	843	82.1	2.62	4.46	10.81
Belief	1121	1149	1067	75.52	10.52	3.35	10.61
Astronomy	1259	1725	1220	84.74	2.15	4.52	8.59

be characterized as strongly comparables. This demonstrates the difficulty of identifying parallel sentences in the extracted comparable corpora. Nevertheless, we can locate comparable sentences and then select subsequently parallel segments. In the following sub-sections, we will focus only on the evaluation of the media domain corpus due to lack of space.

6.2 Comparable Sentences Evaluation

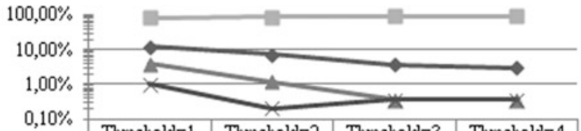
In order to evaluate the effectiveness of the combined anchor-point-based method for comparable sentences alignment, we conducted a manual evaluation of the aligned sentence pairs. This evaluation is based on 1000 pairs of automatically selected comparable sentences randomly chosen. It describes the data distribution for different values of the threshold "Anch". Figure 2 characterizes the aligned sentences (parallel, semi-parallel, comparable and semi-comparable) with different threshold values for the similarity measure.

We define the following terms:

- **Parallel sentences**: pairs of translated sentences.
- **Semi-parallel sentences**: pairs of translated sentences with some insertion or deletion.
- **Comparable sentences**: pairs of non translated sentences but sharing a same topic.
- **Semi-comparable sentences**: pairs of non comparable sentences containing some translated terms.

It is clear that when we increase the threshold the percentage of semi-comparable sentences decreases and the percentage of comparable sentences increases. With Anch = 2, the percentage of parallel and semi-parallel sentences decreases. This is due to some short parallel sentences (e.g. Titles of documents) containing only one anchor point.

In order to maintain a maximum coverage value, we chose a threshold value equal to one for the rest of evaluations.



	Threshold=1	Threshold=2	Threshold=3	Threshold=4
◆ Semi-comparable	12,39%	7,40%	3,82%	3,04%
■ Comparable	82,64%	91,18%	95,45%	96,24%
▲ Semi-parallel	3,96%	1,22%	0,36%	0,36%
× Parallel	1,02%	0,20%	0,36%	0,36%

Fig. 2 Evaluation of comparable sentences alignment process

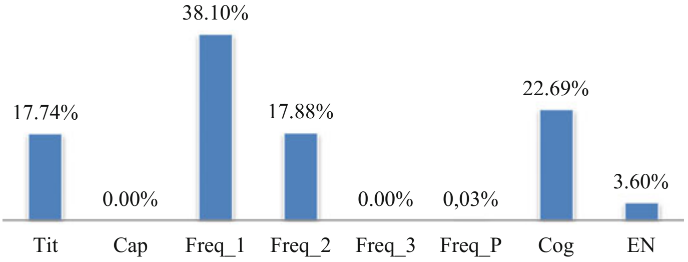


Fig. 3 Anchor point distribution in comparable sentences

Figure 3 shows the distribution of different anchor point types (title, caption, words that occur only once, words that occur twice, words that occur three times, most frequent words, named entities and cognates) in aligned comparable sentences. We notice that Freq_3 and Caption do not match any pair of sentences, whereas Freq_1 (words that occur only once) is the most used feature.

An error analyses of the detected sentences pairs shows that errors are mainly caused by:

- **Wrong lemmatization of rich morphological words**. An example is the anchor point "نرد/nrd[1] - dé (dice)". In this example, the Arabic word "نرد" in the non-vocalized form have two potential tokenizations: the first one is "رد + ن" (we respond). In this case, the lemma is "رد". The second one is "نرد" (dice). In this case, the lemma is "نرد". MADA toolkit which we use for lemmatization returns wrongly the lemma "رد". Then two non-comparable sentences are aligned; the first one contains the lemma رد"" and the second sentence contains the lemma "dé/dice".

- **Ambiguity of non-vocalized arabic words**. For example, the Arabic word "أم/ Om" has two morphological analyses: which refers to the disjunction "or" and which refers to the noun "mother" in English. Our system wrongly matches the anchor point "أم"-"mère/mother" to a pair of non-comparable sentences; the French one contains the word "mère/mother" and the Arabic one contains the preposition "أم/or".

- **Date and number anchor points can wrongly accept non-comparable sentences**. The same numeral characters which appear in a pair of non-comparable sentences are considered as cognates.

- **Erroneous cognates**. For example, the Arabic-French pair of "معبر/mEbr"-"membre". In this example, the Arabic word means "a passageway" whereas the

[1] All Arabic transliterations are provided using the Buckwalter transliteration scheme.

French word means "a member". These words look as cognates but they are semantically different.

- **Translation errors of hapax and words that occur twice in a document**. An example is the word "1er" (1st) which is considered as hapax in a document. This word is translated into "1" whereas "1" may be used in other contexts.

6.3 Parallel Phrases Evaluation

We have extracted 31 842 translations (in media domain). In order to get an idea about the extracted data quality, we took randomly 1000 pairs of fragments and classified them into three classes:

- **Parallel**: perfect translation.
- **Comparable**: there are some insertion, suppression or relations of hyponymy or hyperonymy.
- **Non comparable**: phrases are independents.

Table 2 shows the distribution of the extracted translations by degree of parallelism, in both cases of translation filter:

- **−Tol**: select only pairs of sequences that co-occur exactly in a comparable sentences.
- **+Tol**: accept pairs of translated sequences containing insertions or deletions.

We note that taking into account a certain tolerance in translation filter improves the value of precision. This makes it possible to improve the percentage of parallel segments by 5%, decrease the percentages of comparable segments by 1% and the rate of non-comparable segments by 4%.

Our principle purpose is to use the extracted data for the adaptation of a generic machine translation to a specific domain. We used phrase-based SMT systems trained with the Moses toolkit [14]. Word alignment is done with GIZA++ [19]. We implemented a 5-gram language model using the SRILM toolkit [36]. We tokenized the Arabic side of the training, development and test data using the MADA + TOKAN morphological disambiguation system [26]. French preprocessing of the training, tuning and test data simply included down-casing and separating punctuation from words.

A summary of the size of the used data sets is given in Table 3. Our domain-generic parallel corpora are composed of many parallel data: sentence aligned multiUN contains 2 769 361 pairs of sentences, news-commentary contains

Table 2 Distribution of translation hypotheses

	Parallel (%)	Comparable (%)	Non comparable (%)
+Tol	49.2	12.4	38.4
−Tol	54.31	11.42	34.26

Table 3 Sizes of Arabic–French data sets

Corpus	Nb of sentences	Nb of tokens (ar–fr)
Specific tuning corpus	0.6 K	8.2 K–8.4 K
Specific test corpus	0.4 K	3.5 K–3.4 K
Specific language modeling corpus	22.5 K	61.3 K
Specific parallel corpus based on compositional approach +Tol	31.8 K (fragments)	48.3 K–43.5 K
Specific parallel corpus based on compositional approach −Tol	26.8 K (fragments)	39.4 K–36.2 K
Specific parallel corpus based on [37] approach	15.89 K	348.5 K–90.3 K

90 753 pairs of sentences, nist08 contains 813 pairs of sentences and 15 500 pairs of NEs and a bilingual lexicon composed of 235 938 pairs of terms. The French side of these general-domain parallel corpora with the French Euronews corpus are used for general-domain language modeling. The generic-domain tuning data is the test data of the first edition of TRAD 2012. It is composed of different issues of the Arabic newspaper "Le Monde Diplomatique". It contains 423 pairs of sentences. Because of the lack of a media domain parallel corpus, we constructed manually our domain-specific parallel corpora for tuning and testing as follows: after mining comparable sentences from the media domain comparable corpus, the rest of sentences are used for tuning and testing. We select randomly 600 sentences for tuning and 400 sentences for testing from the Arabic side. We used human translators to translate these Arabic sentences into French. In this way we guarantee that tuning and test data are totally different from training data as they are manually translated. Furthermore, the cosine similarity is calculated to verify the distance between tuning and test data, we obtained 0.025 of similarity. The domain-specific monolingual corpus is extracted from the French Wikipedia articles in media domain. This corpus is constructed automatically using the method of domain-specific comparable corpora building, except the constraint of inter-language link to an Arabic article.

In addition, we implement a hybrid length and lexical based approach [35] to detect parallel sentences from comparable one. Our baseline system is a domain-generic SMT, trained with the domain-generic data described in Table 3. Several experiments are done in order to adapt the domain-generic SMT to the media domain.

- **Baseline**: trained with domain-generic training and tuning data described in Table 3.
- **Dev-sp**: trained with domain-generic training data and domain-specific tuning data.
- **Dev + LM-sp**: trained with the same translation model and tuning data of Dev-sp and use a domain-specific corpus in the target language added to the domain-generic data for language modeling.

- **Dev + LM + TMC + Tol**: trained with the same tuning data and language model as Dev + LM-sp system and use parallel phrases based on compositional +Tol based approach and the domain-generic parallel corpus to train the translation model.
- **Dev + LM + TMC − Tol**: trained with the same tuning data and language model as Dev + LM-sp system and use parallel phrases based on compositional −Tol based approach and the domain-generic parallel corpus to train the translation model.
- **Dev + LM + TMS**: trained with the same tuning data and language model as Dev + LM-sp system and use parallel sentences based on the [37] approach and the domain-generic parallel corpus to train the translation model.
- **Dev + LM + TMC + S**: trained with the same tuning data and language model of Dev + LM-sp system and combine parallel phrases based on compositional-based approach, parallel sentences based on the [37] approach and the domain-generic parallel corpus for translation modeling.

We should note that language model adaptation consists of adding the specialized data to the initial general data. A linear or log-linear interpolation of the two language models are impossible due to the reduced size of the domain-specific data.

Considering the domain-generic and domain-specific corpora, different adaptation strategies of the translation models are explored.

(a) Concatenation of new data to the initial generic data to construct a single translation model.
(b) Linear interpolation with adopting the same weights for the two models [33].
(c) Linear interpolation by favoring the specific translation model against the generic model [33].

Translation results obtained on the Specific test set are reported in terms of BLEU and OOV scores in Table 4.

Table 4 shows that the best results are obtained with a domain-specific tuning data and domain-specific translation and language models adapted with the strategy (a).

When integrating a domain-specific tuning data and a domain-specific language model, we observed a relative improvement of 5.56% of the BLUE score and 1.6 points of the OOV score compared to the basic system (Baseline). Furthermore, the Dev + LM + TM + Tol system reaches a BLUE score of 33.87%, when using the specialized parallel corpus based on the compositional approach (+Tol) concatenated to the domain-generic data; which introduces a relative improvement of 10.72% of the BLUE score compared to the baseline system and an improvement of 4.9% in the BLUE score compared to the DEV + LM-sp system. Thus, the Dev + LM + TM + Tol system is considered as the best system in terms of BLUE score.

Although the extracted data is not very large, the percentage of out of vocabulary words decreases by 5.39 points when integrating these specialized data into the translation model. This percentage decreases further when using of the data based on the hybrid approach [37]. This is due to the large coverage of this corpus which

Table 4 SMT results for Arabic to French

Combinaision	Adaptation strategy	% BLEU	% OOV
Baseline	–	30.59	9.89
Dev-sp	–	31.12	9.43
Dev + LM-sp	–	32.29	8.23
Dev + LM + TM-Tol	(a)	33.01	3.72
Dev + LM + TM + Tol	(a)	33.87	2.84
	(b)	30.17	4.11
	(c)	25.33	3.63
Dev + LM + TMS	(a)	32.18	1.3
	(b)	29.36	4.68
	(c)	21.21	2.67
Dev + LM + TMS + Tol	(a)	33.30	1.84
	(b)	30.28	4.68
	(c)	27.5	4.68

Table 5 Significance of SMT improvements in terms of P-Value

Systems	Blue	P-Value	Nist	P-Value
Dev + LM-sp	32.29		6.57	
Dev + LM + TM − Tol	33.01	0.15	6.57	0.41
Dev + LM + TM + Tol	33.87	0.02	6.71	0.04
Dev + LM + TMS + Tol	33.3	0.08	6.50	0.16

is larger than the corpus based on our compositional approach. Note that this improvement in OOV score was not accompanied by an improvement in BLUE score. This demonstrates the noisy of the data based on the hybrid approach. Fusion of the two corpus improves the results of the Dev + LM + TMS system in terms of BLUE score without reaching the Dev + LM + TM + Tol system BLEU score.

Table 5 evaluates the significance of the improvement obtained with Dev + LM + TM−Tol, Dev + LM + TM + Tol and Dev + LM + TMS + Tol systems against the Dev + LM-sp system. The most statistically significant improvement is obtained with the Dev + LM + TM + Tol system, in terms of BLUE and NIST score (P-Value < 0.05).

A first analysis of the Out Of Vocabulary words of the Dev + LM + TM + Tol system showed that 26.4% of these words were not translated due to tokenization errors (pre-processing of the corpus of text). Most specialized terms attached to a punctuation mark (e.g. (mbc), DRAMA., HD.) are not recognized by the tokenization process and subsequently are not translated. Manual tokenization of OOV words from the test corpus before the decoding process improves the blue score by 0.18 points and the OOV score by 0.28. In a second analysis of OOV words, we found that 32% of these words are written in foreign languages most of which are in English (e.g. Broadcasting, Sylvanas).

Table 6 shows an example of an Arabic sentence with the French reference, taken from the test corpus, in addition to different translations produced by various

Table 6 Example of translations of an Arabic sentence produced by various implemented systems

Arabic sentence	إس إم إنترتينمنت هي شركة تسجيلات كورية مستقلة ووكالة مواهب ومنتج وناشر لموسيقى البوب.
Buckwalter translitteration	As Am Antrtynmnt hy $rkp tsjylAt kwryp mstqlp wwkAlp mwAhb wmntj wnA$r lmwsyqy Albwb
Reference	SM Entertainment est une entreprise coréenne d'enregistrements indépendante, agence de talents, producteur et éditeur de la musique pop
Dev + LM + TM − Tol	Las ou enregistrements est une entreprise coréenne indépendant et l'Agence de talents et de producteur et éditeur de musique pop
Dev + LM + TM + Tol	SM Entertainment est une entreprise coréenne enregistrements indépendant et l'Agence de talents et de producteur et éditeur de musique pop
Dev + LM + TMS	Avex Trax coréenne indépendant et l'Agence de talents et productive et éditeur de musique pop
Dev + LM-sp	Wallace mre est une entreprise coréenne enregistrements indépendant et l'Agence de talents et productif et éditeur de musique pop
Dev-sp	Wallace ou des enregistrements est une entreprise coréenne indépendant et l'Agence de talents et de produit et éditeur de musique pop
Baseline	Wallace ou enregistrements est une entreprise coréenne indépendant et l'Agence de talents et productif et éditeur de musique pop

systems we have implemented. In this example, we observe the gradual improvement of translations when adding new specific data. Thus, the Dev + LM + TM + Tol produces the best translation, which is very close to the reference.

7 Conclusion

We have presented a novel model for mining domain-specific parallel data from Wikipedia. This model combine the use of (i) the taxonomy structure of Wikipedia articles to extract domain-specific comparable data, (ii) the concept of anchor points for comparable sentences alignment and (iii) the compositional based approach for parallel data mining. Experimental results, obtained using Arabic and French Wikipedia encyclopedia allow to jointly validate the extraction of domain-specific comparable and parallel corpora and the proposed adaptation methods. The best adapted system, trained on a combination of the baseline and the extracted data, improves the baseline by 3.3 BLEU points. Preliminary experiments with self-training also demonstrate the potential of this technique.

As a follow-up, we intend to investigate the evolution of the translation results as a function of the precision/recall quality of the extracted corpus, and of the quality of the automatically translated data. Furthermore, we plan to address the problem of

Out Of Vocabulary (OOV) words using word embedding. We have also only focused here on the adaptation of the translation model. We expect to achieve further gains when combining these techniques with LM adaptation techniques.

References

1. Aker, A., Feng, Y., Gaizauskas, R.: Automatic bilingual phrase extraction from comparable corpora. In: Proceedings of CICLING 2012, pp. 23–32, Mumbai (2012)
2. Barrón-Cedeño, A., España-Bonet, C., Boldoba, J., Marquez, L.A: Factory of comparable corpora from wikipedia. In: Proceeding of 8th Workshop on Building and Using Comparable Corpora, pp. 3–13, Beijing, China (2015)
3. Benoit, S., Darla, F.: Building a free french wordnet from multilingual resources. In: International Language Resources and Evaluation (LREC'08), Marrakech, Morocco (2008)
4. Brown, P., Pietra, S., Pietra, V., Mercer, R.: The mathematic of statistical machine translation: parameter estimation. Comput. Linguist. **19** (1993)
5. Buckwalter, T.: Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania. Catalog: LDC2002L49 (2002)
6. Chu, C., Nakazawa, T., Kurohashi, S.: Integrated parallel sentence and fragment extraction from comparable corpora: a case study on Chinese–Japanese wikipedia. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **15**(2), 101–1022 (2016)
7. Delpech, E.: Traduction assistée par ordinateur et corpus comparables: contributions la traduction compositionnelle. PhD thesis, Univérsité de Nante, France (2013)
8. Elkateb, S., Black, W., Vossen, P., Farwell, D., Pease, A., Fellbaum, C.: Arabic wordnet and the challenges of arabic. In: The Challenge of Arabic for NLP/MT, pp. 15–24, London (2006)
9. Fung, P., Cheung, P.: Mining very-non-parallel corpora: parallel sentence and lexicon extraction via bootstrapping and em. In: Proceedings of Empirical Methods on Natural Language Processing (EMNLP), pp. 57–63, Barcelona, Spain (2004)
10. Gamallo, O.P., Loopez, I.G.: Wikipedia as multilingual source of comparable corpora. In: Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, pp. 21–25 (2010)
11. Grefenstette, G.: The world wide web as a resource for example-based machine translation tasks. In: ASLIB99 Translating and the Computer, vol. 21 (1999)
12. Haruno, M., Yamazaki, T.: High-performance bilingual text alignment using statistical and dictionary information. In: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL96), pp. 131–138 (1996)
13. Kay, M., Roscheisen, M.: Text-translation alignment. Comput. Linguist. **19**, 121–142 (1988)
14. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07, pp. 177–180. Association for Computational Linguistics, Stroudsburg, PA, USA (2007)
15. Lardilleux, A., Lepage, Y.: Hapax legomena: their contribution in number and efficiency to word alignment. In: Proceedings of Human Language Technology. Challenges of the Information Society, Third Language and Technology Conference. Lecture Notes in Computer Science 5603, pp. 440–450, Poznan, Poland (2007)
16. Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Phys. Dokl. **10** (1966)
17. Li, B., Gaussier, E.: Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In: Proceedings of the International Conference on Computational Linguistics (COLING10), pp. 644–652 (2010)

18. Munteanu, D.S., Marcu, D.: Extracting parallel sub-sentential fragments from non-parallel corpora. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp. 81–88 (2006)
19. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. **29**, pp. 19–51 (2003)
20. Pal, S., Pakray, P., Naskar, S.K.: Automatic building and using parallel resources for smt from comparable corpora. In: Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra), pp. 47–56, Gothenburg, Sweden (2014)
21. Patry, A., Langlais, P.: Paradocs: l'entremetteur de documents paralèlles indpendant de la langue. Traitement Automatique des Langues **51**, 41–63 (2010)
22. Plamada, M., Volk, M.: Towards a wikipedia-extracted alpine corpus. In: The Fifth Workshop on Building and Using Comparable Corpora, Istanbul, Turkey (2012)
23. Prochasson, E., Morin, E., Kageura, K.: Anchor points for bilingual lexicon extraction from small comparable corpora. In: Proceedings of 12th Conference on Machine Translation Summit (MT Summit XII), pp. 284–291, Ottawa, Ontario, Canada (2009)
24. Rapp, R., Sharoff, S., Zweigenbaum, P.: Recent advances in machine translation using comparable corpora. Nat. Lang. Eng. **22**, 4 (2016)
25. Romary, L., Bonhomme, P.: Parallel alignment of structured documents. In: J. Vronis (ed.) Parallel Text Processing, pp. 201–218 (2000)
26. Sadat, F., Habash, N.: Combination of arabic preprocessing schemes for statistical machine. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 1–8 (2006)
27. Samy, D., Moreno, A., Guirao, J.: A proposal for an arabic named entity tagger leveraging a parallel corpus (Spanish–Arabic). In: Proceedings of Recent Advances In Natural Language Processing RANLP, pp. 459–465 (2005)
28. Sellami, R., Deffaf, F., Sadat, F., Hadrich Belguith, L.: Improved statistical machine translation by cross-linguistic projection of named entities recognition and translation. Computacion y Sistemas **19**, 4 (2015)
29. Sellami, R., Sadat, F., Hadrich Belguith, L.: Exploiting multiple resources for japanese to english patent translation. In: Proceedings of MT Summit XI Workshop on Patent Translation, pp. 34–39, Nice, France (2013)
30. Sellami, R., Sadat, F., Hadrich Belguith, L.: Traduction automatique statistique partir de corpus comparable: application au couple de langues arabe-français. In: Proceedings of CORIA 2013, pp. 431–440, Neuchtel, Switzerland (2013)
31. Sellami, R., Sadat, F., Hadrich Belguith, L.: Improving named entity translation by exploiting noisy parallel corpora. In: Izwaini, S. (ed.) Paper in Translation Studies, Newcastle upon Tyn, pp. 179–198, Cambridge Scholars Publishing (2015)
32. Semmar, N., Saadane, H.: Etude de l'impact de la translittération de noms propres sur la qualité de l'alignement de mots partir de corpus parallèles franais-arabe. In: Actes de la 21e confrence sur le Traitement Automatique des Langues Naturelles, pp. 268–279, Marseille, France (2014)
33. Sennrich, R., Schwenk, H., Aransa, W.: A multi-domain: translation model framework for statistical machine translation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria (2012)
34. Skadia, I.: Analysis and evaluation of comparable corpora for under-resourced areas of machine translation. In: The 5th Workshop on Building and Using Comparable Corpora, LREC 2012, pp. 17–19 (2012)
35. Smith, J.R., Quirk, C., Kristina, T.: Extracting parallel sentences from comparable corpora using document level alignment. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter conference of the Association for Computational Linguistics, pp. 403–411, Los Angeles, California (2010)
36. Stolcke, A.: Srilman extensible language modeling toolkit. In: Proceeding of ICSLP, pp. 901–904, Denver (2002)

37. Varga, D., Lszl, N., Peter, H., Andrs, K., Viktor, T., Viktor, N.: Parallel corpora for medium density languages. In: Proceedings of the RANLP, pp. 590–596 (2005)
38. Wolk, K., Rejmund, E., Marasek, K.: Multi-domaIn: machine translation enhancements by parallel data extraction from comparable corpora (2016). arXiv:1603.06785
39. Zesch, T., Gurevych, I., Mhlhuser, M.: Comparing wikipedia and german word-net by evaluating semantic relatedness on multiple datasets. In Proceedings of NAACL, pp. 205–208 (2007)
40. Zobel, J., Dart, P.: Phonetic string matching: lessons from information retrieval. In: Proceedings of the Eighteenth ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 166–173, Zurich, Switzerland (1996)