Using Analogical Associations to Acquire Chinese-Japanese Quasi-parallel Sentences

Wei Yang† **Hao Wang**‡ **Yves Lepage***
Graduate School of IPS, Waseda University, 2-7 Hibikino, Wakamatsu
Kitakyushu Fukuoka, 808-0135, Japan
†kevinyoogi@akane.waseda.jp
‡oko_ips@ruri.waseda.jp
*yves.lepage@waseda.jp

Abstract

Bilingual parallel corpora are an extremely important resource in current data-driven Natural Language Processing (NLP) systems, especially Machine Translation (MT). There exist numerous freely available bilingual or multilingual parallel corpora for language pairs that involve English, but almost none between Chinese and Japanese. We show how to construct a free Chinese-Japanese quasi-parallel corpus by using analogical associations based on sentential resources collected from the Web. We first over-generate new candidate sentences by analogy. Then, so as to ensure fluency of expression and adequacy of meaning, we filter them by attested N-sequences and obtain valid new sentences at least 99% correct on the grammatical level. Finally, we deduce translation relations across languages based on similarity computation and obtain thousands of quasi-parallel Chinese-Japanese sentence pairs with their associated similarity scores from several tens of thousands sentences without copyright problems as all sentences have been created by our method.

Keywords: Quasi-parallel Corpus, Analogies, Clustering, Sentences Generation, Machine Translation

1 Introduction and Motivation

Corpus-based techniques to statistical or example-based machine translation demand large parallel corpora in multi-domain with proper quality as training data. There already exist freely available corpora for European languages, such as the Europarl parallel corpus [1] or the JRC-Acquis corpus (more than 20 European languages) [2], but currently almost none between Chinese and Japanese publicly available parallel corpora for users and researchers.

Some research institutions have tried to construct Chinese-Japanese bilingual parallel corpora, e.g., NICT (National Institute of Information and Communications Technology, Japan.) or Kyoto U. (Kyoto University, Japan.):

- NICT created a Japanese-Chinese corpus of 38,383 sentences by selecting Japanese sentences from the Mainichi Newspaper and translating them manually into Chinese. They then annotated the corpus with morphological and syntactic structures and alignments at word and phrase levels [3].
- Kyoto U. Kurohashi-Kawahara Lab[1] created the Japanese-English-Chinese (JEC) Basic Sentence Data based on the Japanese Basic Sentence Data, automatically extracted from the Kyoto University Case Frame data. Their data contain manually modified 5,304 short sentences, and then manually translated data from Japanese into English and Chinese as a NICT MASTAR Project in Multilingual Translation Laboratory[2].

Such Chinese-Japanese corpora are translated from one language into another language manually. None of them have been constructed automatically. Except for the JEC Basic Sentence Data all the rest is not publicly or freely available, due to copyright problems. These parallel corpora are small in comparison to the above-mentioned multilingual corpora in European languages.

[1] Kurohashi-Kawahara Lab: http://nlp.ist.i.kyoto-u.ac.jp

[2] Multilingual Translation Laboratory: http://www.nict.go.jp/en/univ-com/multi_trans/

The constitution of large collections of aligned sentences is a problem for less documented language pairs. But, it is to be noticed that a less documented language pair may involve two well-documented languages, as is the case for the languages we address here: Chinese and Japanese.

Corpus-based analogical techniques have been previously proposed for machine translation [4], morphology [5], and semantic relations [6]. Recognizing associations between words or sentences is an important task in Natural Language Processing. Such associations can be obtained using analogical relations [hand : glove :: foot : shoe] or [to create : creator :: to translate : translator]. In this research, we propose to construct a quasi-parallel Chinese-Japanese corpus by making use of analogical associations based on Chinese and Japanese linguistic resources collected from the Web. A "*quasi-parallel corpus*" contains sentences that are nearly the exact translation to each other. On the difference with a parallel corpus, the quasi-parallel corpus that we create gives similarity scores between the sentences in the corpus.

We propose to use analogical associations between sentences to cluster large amounts of short sentences collected in both Chinese and Japanese independently. Such clusters can be considered as rewriting models that can generate new sentences. To solve the problem of over-generation, we use N-sequences to filter out dubious newly generated sentences and enforce fluency of expression and adequacy of meaning. Based on the similarity between the clusters across languages, and the similarity between sentences for new sentences generation, we can assess the strength of translation relations between newly created sentences. The set of such newly created sentences, with their translation scores, will constitute a quasi-parallel Chinese-Japanese corpus without copyright problems, as all the sentences contained will have been created by our programs.

2 The Chinese and Japanese Linguistic Resources Collected from the Web

We collect Chinese and Japanese short sentences (less than 30 characters in size) from the Web as our basic experimental data, using an in-house Web-crawler. The use of the Web ensures that our data is made of natural sentences.

2.1 Chinese and Japanese monolingual sentences

The main websites from which we collected monolingual sentences are "Yahoo China[3]", "Yahoo China News[4]", "douban[5]" for Chinese and "Yahoo! JAPAN[6]", "Mainichi Japan[7]", "Rakuten Japan[8]" for Japanese.

Figure 1 illustrates the increasing tendency of the raw-filtered Chinese and Japanese short sentences in one year. We cleaned up these data by filtering out any sentence containing undesirable characters or symbols. For Chinese data, we retained sentences that contain only simplified Chinese characters.

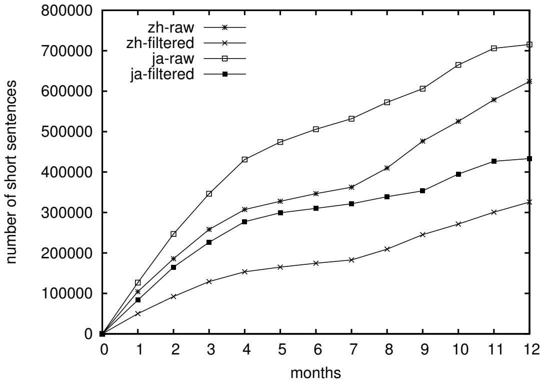

Figure 1. Statistics about Chinese and Japanese raw-filtered short sentences from the Web.

Table 1 shows the statistics about these filtered Chinese and Japanese short sentences. About half (52%) of the Chinese sentences and more than half (60.6%) of Japanese short sentences are kept after filtering. The quality of these kept short sentences has been estimated by hand and is at least 98% correct sentences (p-value = 0.02).

2.2 Chinese-Japanese parallel sentences

We also collected and processed some Chinese-Japanese parallel data, as a part of our experimental data for assessment purposes:

- the JEC Basic Sentence Data (Kyoto U. and

[3] Yahoo China: `http://cn.yahoo.com`

[4] Yahoo China News: `http://news.cn.yahoo.com`

[5] douban: `http://www.douban.com`

[6] Yahoo! JAPAN: `http://yahoo.co.jp`

[7] Mainichi Japan: `http://mainichi.jp`

[8] Rakuten Japan: `http://www.rakuten.co.jp`

Table 1. Statistics on the filtered Chinese and Japanese monolingual short sentences.

	# of different sentences (collected)	# of different sentences (filtered)	size of sentences in characters			total characters	total words
			mean	±	std.dev.		
Chinese	623,929	325,815	11.29	±	7.24	3,609,708	2,445,764
Japanese	715,432	433,292	16.06	±	7.43	7,053,924	4,116,804

Table 2. Statistics on the Chinese-Japanese parallel sentences data.

	Language	# of different sentences	size of sentences in characters			total characters	total words
			mean	±	std.dev.		
JEC	Chinese	5,299	12.31	±	4.40	65,219	43,761
	Japanese	5,304	16.29	±	5.38	86,409	53,654
日语学习网	Chinese	10,960	11.41	±	4.49	185,537	129,003
	Japanese	14,775	17.28	±	8.42	347,055	191,615

NICT, 2011) with 5,304 Chinese-Japanese sentence pairs.

- Chinese-Japanese Learning Corpus: from "日语学习网[9]" and "沪江网[10]", we obtained about 15,302 Chinese-Japanese parallel sentences after processing.

The total number of these Chinese-Japanese parallel corpora in lines is 20,606, including 16,259 different sentences for Chinese and 20,079 different sentences for Japanese. Table 2 given the statistics on these Chinese-Japanese parallel sentences data we used. As an ideal configuration, we suppose the similarity between each sentence pair is 1.000.

3 Building Analogical Clusters

3.1 Proportional analogies

Proportional analogies establish a general relationship between four objects, A, B, C and D. An analogy $A : B :: C : D$ states that 'A is to B as C is to D'. Previous research by Lepage (1998) [7] proposes an efficient algorithm for the resolution of analogical equations. The algorithm is based on counting numbers of occurrences of characters and the computing edit distances between strings of characters. It is given by Formula (1).

$$A : B :: C : D \Rightarrow \begin{cases} |A|_a - |B|_a = |C|_a - |D|_a, \forall a \\ d(A,B) = d(C,D) \\ d(A,C) = d(B,D) \end{cases} \quad (1)$$

where $|A|_a$ stands for the number of occurrences of character a in string A. $d(A, B)$ stands for the edit distance between strings A and B with only insertion and deletion as edit operations. It can be computed very fast using the fast bit string algorithm described in (Allison and Dix, 1986) [8].

In our research, we extract pairs of sentences that follow the above formula for proportional analogies. For instance, the two following pairs of Japanese sentences are said to form an analogy:

紅茶が飲みたい。 : あなたは紅茶が好きですか。 :: ビールが飲みたい。 : あなたはビールが好きですか。

I'd like a black tea. : *Do you like black tea?* :: *I'd like a beer.* : *Do you like beer?*

Because the relational similarity between the sentence pair on the left side of '::' is the same as between the sentence pair on the right side. We call any such two pairs of sentences sentential analogies.

When several sentential analogies involve the same pairs of sentences, they form a series of

[9] 日语学习网 (Japanese Learning net): `http://jp.tingroom.com`

[10] 沪江网 (HuJiang): `http://www.hujiang.com`

analogous sentences, and they can be written on a line like in:

$$A : B :: C : D :: E : F :: \ldots$$

More conveniently, they can also be written on a sequence of several lines like:

$$\begin{array}{c} A : B \\ C : D \\ E : F \\ \ldots : \ldots \end{array}$$

We shall call analogical cluster for such a sequence of lines, where each line contains one sentence pair and where any two pairs of sentences form a sentential analogy. The size of a cluster is the number of its sentential pairs. The following example shows three possible sentential analogies and the size of the cluster is 3.

紅茶が飲みたい。	:	あなたは紅茶が好きですか。
ビールが飲みたい。	:	あなたはビールが好きですか。
ジュースが飲みたい。	:	あなたはジュースが好きですか。

3.2 Experiments on clusters production

We performed experiments based on proportional analogy with Chinese and Japanese monolingual data respectively. The number of lines of unique sentences used is 47,674 for Chinese and 95,130 for Japanese (including 16,259 and 20,079 unique Chinese and Japanese sentences of 20,606 Chinese-Japanese parallel sentences) from raw-filtered data.

Table 3 summarizes some statistics on the clusters produced. The clusters containing only two pairs of sentences are called small clusters. The others are called large clusters. The larger a cluster, the more productive it is. We rank the clusters by decreasing order of size, and assign an identifier to each cluster accordingly.

Table 3. Statistics on the Chinese and Japanese clusters produced from our data.

	Chinese	Japanese
# of different sentences	47,674	95,130
# of clusters	28,455	37,185
# of small clusters	2,893	8,445
# of large clusters	25,562	28,740
Time spent (h)	5.37	21.34

Table 4 and Table 5 are examples of clusters' obtained. Because different clusters illustrate different linguistic features, the same sentence may be belong to several clusters. As Table 4 and Table 5 show, the Chinese sentence 画面也很漂亮 /huàmiàn yě hěnpiàoliàng/ 'Frames are also very beautiful.' appears in the cluster in Table 4 (indicated with a '▲'). This cluster shows the insertion of the adverbial "也" /yě/ 'also'. The same sentence also appears in the cluster in Table 5. The linguistic interpretation of this cluster is that the substitution of the Chinese noun "画面" /huàmiàn/ 'frames' and adjective "漂亮" /piàoliàng/ 'beautiful' with another two words "使用" /shǐyòng/ 'use' and "方便" /fāngbiàn/ 'convenient' can happen in similar situational or structural contexts. In such situations, left and right sentences are not paraphrases. It shown changes of semantic features.

Table 4. A cluster (identifier 3081) that illustrates the possible insertion of the adverbial "也" /yě/ 'also' in Chinese.

画面可爱	:	画面也可爱
画面精致	:	画面也精致
画面很漂亮	:	▲画面也很漂亮
画面不错	:	画面也不错
画面很不错	:	画面也很不错

Table 5. A cluster (identifier 4215) that illustrates the substitution of the Chinese noun "画面" /huàmiàn/ 'frames' and adjective "漂亮" /piàoliàng/ 'beautiful' with another two words "使用" /shǐyòng/ 'use' and "方便" /fāngbiàn/ 'convenient'.

▲画面也很漂亮	:	使用也很方便
画面非常的漂亮	:	使用非常的方便
画面很漂亮	:	使用很方便
画面漂亮	:	使用方便

4 Generation of New Sentences Using Analogical Associations

4.1 Generation of new sentences

We now show how to generate new sentences based on analogical relations. Following Saussure [9], we use analogy as a synchronic operation by which, given two related forms and only one form, the fourth missing form is coined. Applied on sentences, this principle can be illustrated as follows:

紅茶が飲みたい。 : あなたは紅茶が好きですか。 :: ビールが飲みたい。 : x

⇒ $x =$ あなたはビールが好きですか。

If the objects A, B, C are given, we may obtain an other unknown object D according to the analogical equation $A : B :: C : D$. In this example, the solution of the analogical equation is D = "あなたはビールが好きですか。" (Do you like beer?). If we regard each sentence pair in a cluster as a pair $A : B$ (left to right or right to left), and any short sentence not belonging to the cluster as the object C, the analogical equation $A : B :: C : D$ of unknown D can be forged. Such analogical equations allow us to produce new candidate sentences. Each sentence pair in a cluster is a potential template for the production of new candidate sentences.

4.2 Experiments on new sentences generation and filtering

For new sentences generation, we make use of the clusters we constructed in section 3.2 as rewriting models, and the seed sentences (input data) are the unique Chinese and Japanese short sentences from the 20,606 parallel sentences. In this experiment, we generated new sentences with each pair of sentences in clusters for Chinese and Japanese respectively. This makes it possible to obtain more different and well-formed new sentences. We generated about 62 million Chinese candidate sentences and more than 18 million Japanese candidate sentences. We extracted a sample of 1,000 sentences and checked their quality manually. The quality lies around 19% for Chinese and 50% for Japanese of correct sentences in syntax and meaning. Table 6 details the figures for this experiment.

To filter out invalid and grammatically incorrect sentences and keep only well-formed sentences with high fluency of expression and adequacy of meaning, we eliminate any sentence that contains an N-sequence of a given length unseen in the reference corpus. This technique to assess the quality of outputs of NLP systems has been used in works by C-Y. Lin and E. Hovy for summary generation [10], G. Doddington for machine translation [11], and also Y. Lepage and E. Denoual for filtering paraphrases generation [12]. In our experiment, we introduced begin/end markers to make sure that at least the beginning and the end of a sentence is correct. The best quality was obtained for the values N=6 for Chinese and N=7 for Japanese with the size of reference corpus (in lines) given in Table 6. We obtained 4,898 new valid sentences in Chinese and 8,873 new valid sentences in Japanese. This time, the grammatical quality evaluated on a sample of 1,000 sentences, was at least 99% checked by native speakers, and this means that 99% of the Chinese and Japanese sentences may be considered as grammatically correct. This quality is of the same level as the quality of the resources we started from (see section 2.1).

Table 7 gives examples of newly generated sentences and the result of filtering using unseen N-sequences (N=6) for Chinese. The unacceptable new sentences are struck through in the table. For valid sentences, we remember their corresponding seed sentences and the cluster identifiers they were generated by.

5 Deducing Translation Relations and Acquiring Chinese-Japanese Quasi-parallel Sentences

In this paper, we examine an ideal configuration where the Chinese and Japanese seed sentences are parallel. Thus, for our experiment, the translation relations between the seed sentences rely on the 20,606 Chinese-Japanese parallel corpus. Then, we propose to extract corresponding clusters and compute the similarity between them so as to deduce and construct a Chinese-Japanese quasi-parallel sentences between new valid sentences.

5.1 Extracting corresponding clusters by computing similarity

First, we extract the change between left and right sides in each cluster by finding the longest common subsequence (LCS) [13] between each

Table 6. Statistics for new sentences generation in our experiments with Chinese and Japanese data.

		Chinese	Japanese
Initial data	# of seed sentences	16,259	20,079
	# of clusters	28,455	37,185
Generation	# of candidate sentences	61,565,221 Q=19%	18,403,787 Q=50%
Quality assessment	# of lines in references corpus	269,308	336,877
	# of new valid sentences (with begin/end marks)	4,898 N=6, Q=99%	8,873 N=7, Q=99%

Table 7. For each valid sentence, we remember its corresponding seed sentence and the cluster identifier that produced it.

Seed short sentences		Newly generated sentences	Cluster identifier
我也非常喜欢音乐。	:	我也很喜欢音乐。	944
值得下载的游戏	:	值得推荐的游戏	608
值得推荐的软件	:	值得推荐的游戏	97
我最喜欢的这款软件很有意思。	:	我最喜欢的这款游戏很有意思。	97
		~~这个女孩长得。不错~~	
		~~小孩子很有教育意义。~~	
我早饭吃的面包。	:	我早饭吃的日本料理。	135
		~~则改之。无则加勉~~	

sentence pair. Then, we consider the changes between the left (S_{left}) and right (S_{right}) sides in one cluster as two sets. Finally, we perform the segmentation[11] for these changes in sets to obtain minimal sets of changes made up with words or characters.

Finding the corresponding clusters reduces to compute the similarity between two left sets (S_{left}) and two right sets (S_{right}) between Chinese and Japanese clusters. We make use of the EDR dictionary[12] and traditional-simplified Chinese conversion (Unicode Data-Traditional-Simplified-Variant[13]) and a Kanji-Hanzi Conversion Table[14] to translate all Japanese words into Chinese, or convert Japanese characters into simplified Chinese. We calculate the similarity between Chinese and Japanese word sets according to a classical Dice formula:

$$Sim = \frac{2 \times |S_{zh} \cap S_{ja}|}{|S_{zh}| + |S_{ja}|} \qquad (2)$$

S_{zh} and S_{ja} denote the minimal sets of changes across the clusters (both on the left or right) in both languages (after translation and conversion). In our experiment, the formula for computing the similarity between Chinese and Japanese clusters as given in formula (3):

$$Sim = \frac{1}{2}(Sim_{left} + Sim_{right}) \qquad (3)$$

As the example shown in Table 8, for the Chinese and Japanese clusters with the changes of “小说： 电影很好看” and “小説： いい映画”, after segmentation, we obtained the Chinese translation for the Japanese words 映画 /eiga/ with 电影 /diànyǐng/ (they both mean ‘movie’),

[11] Segmentation toolkits: Mecab, Part-of-Speech and Morphological Analyzer: URL: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html for Japanese and Urheen, a Chinese lexical analysis toolkit (National Laboratory of Pattern Recognition, China) for Chinese.

[12] The EDR Electronic Dictionary: National Institute of Information and Communication Technology (NiCT). URL: http://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html

[13] http://www.unicode.org/Public/UNIDATA/

[14] http://www.kishugiken.co.jp/cn/code10d.html

Table 8. Examples of corresponding clusters with high similarity scores in our result.

Changes in Chinese Cluster	**zh ⇐ ja**	**Changes in Japanese Cluster**	**Similarity**
面包 : 日本料理		パン : 日本料理	1.000
喜欢 : 讨厌		好き : 嫌い	1.000
很 : 非常	EDR dictionary	超 : とても	1.000
但是 : ε	+	でも : ε	1.000
照片 : ε	TS conv.	写真 : ε	1.000
ε : 她	+	ε : 彼女	1.000
ε : 非常	kanji-hanzi conv.	ε : 非常に	0.833
小说 : 电影很好看		**小説 : いい映画**	**0.700**
十分 : 非常		ε : とても	0.500

$$\text{Sim} = \frac{1}{2}\left(\frac{2\times|\{\text{小说}\}|}{|\{\text{小说}\}|+|\{\text{小说}\}|}+\frac{2\times|\{\text{电影}\}|}{|\{\text{电影},\text{很},\text{好看}\}|+|\{\text{いい},\text{电影}\}|}\right)=\frac{1}{2}\left(1+\frac{2}{5}\right)=0.700$$

and 小说 /xiǎoshuō/ with 小説 /shousetsu/ (they both mean 'novel'). Thus, using this method, the similarity cross this two clusters calculated by the formula (3) would be 0.700. We assume that we cannot find Chinese translation of the Japanese word 小説, we also may obtain the same similarity by converting the Japanese kanji 説 /setsu/ in 小説 /shousetsu/ to simplified Chinese hanzi 说 /shuō/.

If we could translate or convert all words and characters in the sets of changes (left and right) for a Japanese cluster into Chinese and match the sets of changes for a Chinese cluster (left and right), the similarity score will be 1.000. In our experiment, by setting a threshold for the similarity score with 0.300, we obtained 6,137 corresponding clusters cross languages.

5.2 Experiments and results

The method we presented using analogical associations to acquire Chinese-Japanese quasi-parallel sentences from the data: (i) 20,606 Chinese-Japanese parallel corpus with translation similarity scores presented in section 2.2; (ii) 4,898 valid newly generated Chinese sentences (8,873 for Japanese) with their corresponding seed sentences, identifiers of the clusters that they generated from, the data we obtained in section 4.2; (iii) 6,137 corresponding clusters we obtained cross languages in section 5.1, composed of the sets of changes in both languages with identifiers and the similarity score.

Relying on the similarity between the clusters across languages and the similarity between the seed sentences, we could assess the strength of translation relations between the newly generated sentences. As a final result, 1,837 quasi-parallel sentences with their translation scores were obtained by our method. Among them 1,124 Chinese-Japanese sentence pairs are true translations (61.2% of 1,837 quasi-parallel sentences).

6 Conclusion

We presented a technique which uses analogical associations to construct a free Chinese-Japanese quasi-parallel corpus based on sentences collected from the Web with the concern of avoiding any copyright problem. The result is a resource of pairs of sentences in Chinese and Japanese with associated translation similarity scores.

From 47,674 sentences in Chinese and 95,130 sentences in Japanese, we could construct 28,455 analogical clusters in Chinese and 37,185 in Japanese. These clusters served as rewriting models to generate new sentences. To ensure fluency of expression and adequacy of meaning we filtered the generated sentences by the N-sequence method. 4,898 Chinese new sentences and 8,873 Japanese new sentences were obtained after filtering. Their grammaticality and semantic validity was evaluated by sampling and was found to be of at least 99% for both Chinese and Japanese. Such valid new sentences are not necessarily the paraphrases compare with the seed sentences. We then deduced the translation relations between newly generated short sentences across both languages, relying on the similarity between the seed sentences and the clusters they were generated from.

In this paper, we examined an ideal configuration where the Chinese and Japanese seed sentences were parallel sentences. We automatically obtained 1,837 quasi-parallel new generated sentences without copyright problems as all sentences have been created by our method.

In future work we propose to extract more quasi Chinese-Japanese parallel sentences based on this method from comparable resources by computing the similarity between the seed sentences and make them freely available to the community to foster research in Chinese-Japanese machine translation.

References

[1] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand, 2005.

[2] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, and Dan Tufiş. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147, 2006.

[3] Yujie Zhang, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. Building an annotated japanese-chinese parallel corpus–a part of nict multilingual corpora. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 71–78, 2005.

[4] Yves Lepage and Etienne Denoual. Purest ever example-based machine translation: detailed presentation and assessment. *Machine Translation*, 19:251–282, 2005.

[5] Jean-François Lavallée and Philippe Langlais. Morphological acquisition by formal analogy. In *Morpho Challenge 2009*, Corfu, Greece, oct 2009.

[6] Peter D. Turney and Michael L. Littman. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1–3):251–278, 2005.

[7] Yves Lepage. Solving analogies on words: An algorithm. In *Proceedings of the 36th Annual Conference of the Association Proceedings of the 36th Annual Conference of the Association for Computational Linguistics (COLING-ACL'98)*, pages 728–735, Montréal, August 1998.

[8] Lloyd Allison and Trevor I. Dix. A bit string longest common subsequence algorithm. *Information Processing Letter*, 23:305–310, 1986.

[9] Ferdinand de Saussure. *Cours de linguistique générale*. Payot, Lausanne et Paris, [1ère éd. 1916] edition, 1995.

[10] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL-2003)*, pages 71–78, 2003.

[11] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology Conference (HLT2002)*, pages 128–132, San Diego, CA, USA, 2002. Morgan Kaufmann.

[12] Yves Lepage and Etienne Denoual. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation. In *the 3rd International Workshop on Paraphrasing (IWP2005)*, pages 57–64, 2005.

[13] Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21:168–173, 1974.