Survey on Comparable Corpora until June 2012

Victor Chakraborty

June 27, 2012

Abstract

Here we present a survey of important work done on Comparable Corpora between the period 1995 to 2012. Unlike parallel corpora, which are clearly defined as translated texts, there is a wide variation of non-parallelism in comparable text. Non-parallelism is manifested in terms of differences in author, domain, topics, time period, language. The most common text corpora have non-parallelism in all these dimensions. The higher the degree of non-parallelism, the more challenging is the extraction of bilingual information. Such a corpus is nevertheless a desirable source of bilingual information, especially for new words. In this report we have first classified the research on comparable corpora into various categories. This is followed by detailed literature survey on comparable corpora and comparability metrics. After that we discuss the work related to the enhancement of comparability metrics in corpus. We conclude with the brief summary of this survey on comparable corpora.

1 Classification of Work on Comparable Corpora

We can broadly classify the research on comparable corpora into the following sections.

- Correlation based extraction
- Vector representation
- Classifiers based extraction
- Linguistic knowledge based extraction

Each of these classes are described in the following sections along with the gist of pioneering work in these domain.

2 Correlation based Extraction

Most of the work on comparable corpora is based on correlation between word co-occurrence. They consider the context of a word as a feature to map the source word to the target word. Moreover most of the work based on this idea is focused towards extraction of bilingual lexicons only rather than parallel sentences.

2.1 Context heterogeneity

Fung proposed a novel context heterogeneity similarity measure between words and their translations in helping to compile bilingual lexicon entries from a non-parallel English-Chinese corpus [Fun95]. Context heterogeneity measures how productive the context of a word is in a given domain, independent of its absolute occurrence frequency in the text. Based on this information, one can derive statistics of bilingual word pairs from a non-parallel corpus. These statistics can be used to bootstrap a bilingual dictionary compilation algorithm. Context heterogeneity vector of a word W is an ordered pair (x, y) where:

$$left\ heterogeneity\ x = \frac{a}{c}$$
$$right\ heterogeneity\ y = \frac{a}{c}$$

$a = number\ of\ different\ types\ of\ tokens\ immediately\ preceding\ W\ in\ the\ text$

$b = number\ of\ different\ types\ of\ tokens\ immediately\ following\ W\ in\ the\ text$

$c = number\ of\ occurrences\ of\ W\ in\ the\ text$

The context heterogeneity of any function word, such as *the*, would have x and y values very close to one, since it can be preceded or followed by many different words. On the other hand, the x value of the word *am* is small because it always follows the word *I*.

To measure the distance between two context heterogeneity vectors, simple Euclidean distance is used.

$$E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2)}$$

2.2 CONVEC

Fung presented another method called CONVEC to capture the context information of the word [FY98] [Fun98]. This method is based on similarity measured on TF-IDF. The TFIDF weight (term frequencyinverse document frequency) is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Highly frequent words get low IDF value.
Now one can visualize the context vector for a word to have the dimension of the bilingual dictionary in use. The i^{th} dimension of this vector is $w_i = TF_i * IDF_i$. It is zero if the i^{th} word does not appear in the context of the word. Similarly they obtain the context vectors of all unknown words in the source language and context vectors of all candidate words in the target language. To locate translation candidates for any word, they compared the context vector of that word with context vectors of all words in other language using Cosine Similarity.

2.3 Relation Matrix

Another new concept called Word Relation Matrix is used to find translated pairs of words and terms from non-parallel corpora, across language groups [FM97]. The algorithm is as follows :

1. Given a bilingual list of known translation pairs ($i.e$, seed words)
2. For every unknown word or term e in language 1, find its correlation with every word in the seed word list in language 1 $\Rightarrow$ relation vector $WORM1$
3. Similarly for unknown words c in language 2, find its correlation with every word in the seed word list in language 2 $\Rightarrow$ relation vector $WORM2$
4. Compute $correlation(WORM1, WORM2)$. If it is high, e and c are considered as a translation pair.

2.4 Phrase Frequency based methods

This is again based on the correlation between the co-occurrences of words that are translations of each other This is the only statistical clue used throughout this paper [Rap99]. It is further assumed that there is a small dictionary available at the beginning, and the aim is to expand this base lexicon. Using a corpus of the target language, they first computed the co-occurrence matrix whose rows are all word types occurring in the corpus and whose columns are all target words appearing in the base lexicon. The correlation vector is created from this. Since the base lexicon is small and only some of the translations are known, all unknown words are discarded from the vector and the vector positions are sorted in order to match the vectors of the target-language matrix. With the resulting vector, similarity calculation based on log likelihood is done. For counting word co-occurrences, in most other studies a fixed window size is chosen. However this approach is dropped in this paper. Position of the word with respect to other words is also taken into account.

2.5 Iterative Extraction

This method involves extraction of bilingual lexicons from highly non-parallel corpora [FC04]. This exploits the IBM Model 4 type EM algorithm. First it ranks the documents using similarity measures discussed above. Then it tries to extract parallel words from the 'good' document pairs. But unlike previous methods,they extend this with an iterative bootstrapping framework based on the principle of *find-one-get-more*,which claims that documents found to contain one pair of parallel sentences must contain others even if the documents are judged to be of low similarity.Documents are rematched based on extracted sentence pairs, and mining process is refined iteratively until convergence. This novel *find-one-get-more* principle allows one to add more parallel sentences from dissimilar documents.

Their algorithm can be outlined as follows:

1. Document preprocessing
2. Initial document matching
3. Sentence matching
4. EM lexical learning from matched sentence pairs.
5. Document rematching and *find-one-get-more*
6. Convergence.

Word correlations are computed from general likelihood scores based on the co-occurrence of words in common segments. Segments are either sentences, paragraphs, or string groups delimited by anchor paints:

$$Pr(w_s = 1) = \frac{a+b}{a+b+c+d}$$
$$Pr(w_t = 1) = \frac{a+c}{a+b+c+d}$$
$$Pr(w_s = 1, w_t = 1) = \frac{a}{a+b+c+d}$$

$$a = number\ of\ segments\ where\ both\ words\ occur$$
$$b = numberof\ segments\ where\ onlyw_s\ occur$$
$$c = numberof\ segments\ where\ only\ w_t\ occur$$
$$d = numberof\ segments\ where\ neither\ words\ occur$$

This method is the modified version of Context heterogeneity similarity matrix suggested by Fung 1995 paper.

2.6 Combiniation of Context and Lexical Information

This work combines various models to get better result [DGS02]. They have used the context vector as described before as basic building block. They have introduced the use of multilingual thesaurus for lexical translation. They have calculated the translation probabilities using this thesaurus and used this as weighted edge between source and translation. These models are optimally combined to produce results which are 30% more accurate than standard results.

2.7 Domain Specific Bilingual Dictionary Extraction

Domain specific method for extraction of bilingual lexicon from Medical corpus [CZ02]. Context vectors are generated for each word in the source language on a window length 7. These context vectors are then translated into target language using a small bilingual dictionary. This translated vector is then compared with all possible context vectors in target language with Jaccard and Cosine similarity.

$$Jaccard(V,W) = \frac{\sum_k v_k w_k}{\sum_k v_k^2 + \sum_l w_l^2 - \sum_m v_m w_m}$$

$$Cosine(V,W) = \frac{\sum_k v_k w_k}{\sqrt{\sum_k v_k^2}\sqrt{\sum_l w_l^2}}$$

3 Vector Representation

These approaches are based on treating sentences as vectors and using information retrieval algorithms to extract them. Feature other than context are also considered here. Each sentence is treated as a vector in a feature space. Similarity measures are then used to identify close sentence pairs.

3.1 Geometric Interpretation of Bilingual Text

Gaussier et al. [GRM+04] presented a geometric view on bilingual lexicon extraction from comparable corpora, which allows to re-interpret the methods proposed so far and identify unresolved problems. This motivates three new methods that aim at solving these problems. Empirical evaluation shows the strengths and weaknesses of these methods, as well as a significant gain in the accuracy of extracted lexicons.

They denote by $s_i, 1 \leq i \leq p$ and $t_j, 1 \leq j \leq q$ the source and target words in the bilingual dictionary D. D is a set of n translation pairs (s_i, t_j), and may be represented as a pXq matrix M, such that $M_{ij} = 1\ iff\ (s_i, t_j) \epsilon D$(and 0 otherwise).
One can assume that there are m distinct source words $e_1, , e_m$ and r distinct target words $f_1, , f_r$ in the corpus. The association measure $a(v, e)$ may be viewed as the coordinates of the m-dimensional context vector $\vec{v}$ in the vector space formed by the orthogonal basis $(e_1, , e_m)$.
The similarity of the vectors are calculated as the dot product between $\vec{v}$ and the translation of $\vec{w}$.

$$\left\langle \vec{v}, tr(\vec{w}) \right\rangle = \sum_{(e,f)\epsilon D} a(v,e)a(w,f)$$

This approach solves the problem of polysemy/synonymy and coverage.

3.2 Information Retrieval based Extraction

Extraction of parallel corpus may be compared with document retrieval problem in IR [SN04]. The context of a word is viewed as query. The context of each candidate translation is viewed as document. They employed the language modeling approach for the retrieval problem. In this approach a language model is derived from each document D. Then the probability of generating the query Q according to that language model, $P(Q|D)$, is estimated. The document with the highest $P(Q|D)$ is the one that best matches the query. The language modeling approach to IR has been shown to give superior retrieval performance compared with traditional vector space model

3.3 Text Extraction based on Signal Processing

The work done here for extracting parallel fragments is inspired from signal processing approach [MM06]. First documents are matched for probable good pairs. All sentence pairs from these document pairs are taken and they are passed through the second step in the pipeline,the candidate selection filter. This step discards pairs which have very few words that are translations of each other. To all remaining sentence pairs the fragment detection method is applied. The approach is to consider the target sentence as a numeric signal, where translated words correspond to positive values , and the others to negative ones. We want to retain the parts of the sentence where the signal is mostly positive. This can be achieved by applying a smoothing filter to the signal, and selecting those fragments of the sentence for which the corresponding filtered values are positive. The values for translation can be calculated from the probabilities and count from the corpus.

4 Classifiers based Extraction

This method though almost identical with the previous one, uses classifiers to separate the good sentence(or phrase) pairs from the bad one. Feature generation is the crux of these approaches.

4.1 Maximum Entropy Model

Here a Maximum Entropy classifier is used to classify good pair of sentences from bad pair [MFM04]. The resources used in this are a dictionary and small amount of parallel data for training. The feature used is only the translation count. The ME principle suggest that the optimal parametric form of the model of data , taking into account the constraints imposed by the feature functions is a log linear combination of these functions. The resulting model has free parameters , the features weights. The parameter values that maximize the likelihood of a given training corpus can be computed using algorithms like GIS or its improved version IIS. The feature function used here takes into account the explicit word alignment of in the sentence. Other features considered are :

1. Length of the sentences.

2. Percentage of the words with translations from both side.

3. Percentage of words with no translations

4. Alignment score

5. Length of the longest span

They have shown results in precision recall as well as in BLEU.

4.2 Support Vector Machine

In this paper, the authors describe the use of annotated datasets and Support Vector Machines to induce larger monolingual para-phrase corpora from a comparable corpus of news clusters found on the World Wide Web [BD05]. Features include: morphological variants; WordNet synonyms and hypernyms; log-likelihood-based word pairings dynamically obtained from baseline sentence alignments; and formal string features such as word-based edit distance. Use of this technique dramatically reduces the Alignment Error Rate of the extracted corpora over heuristic methods based on position of the sentences in the text. The main feature classes were:

1. String Similarity Features: All sentence pairs were assigned string-based features, including absolute and relative length in words, number of shared words, word-based edit distance, and lexical distance, as measured by converting the sentences into alphabetized strings of unique words and applying word based edit distance.

2. Morphological Variants: Another class of features was co-occurrence of morphological variants in sentence pairs. Sentences were stemmed using a rule-based stemmer, to yield a lexicons of morphologically variant word pairs. Each word pair was treated as a feature.

3. WordNet Lexical Mappings: Synonyms and hypernyms were extracted from WordNet using the morphological variant lexicons from the initial sentences as keywords. The theory here is that as additional paraphrase pairs are identified by the classifier, new information will be added thereby augmenting the range of paraphrases available to be learned.

4. Word Association Pairs: To augment the above resources, they dynamically extracted from the corpus, possibly-synonymous word pairs using a log-likelihood algorithm for machine translation. To minimize the damping effect of the overwhelming number of identical words, these were deleted from each sentence pair prior to processing. The algorithm was then run on the non-identical residue as if it were a bilingual parallel corpus.

5. Composite Features: From each of the lexical feature classes, they derived a set of more abstract features that summarized the frequency with which each feature or

class of features occurred in the training data, both independently, and in correlation with others. These had the effect of performing normalization for sentence length and other factors.

5 Linguistic Knowledge Based Extraction

Linguistic knowledge is used to prepare templates. These templates are then used as filters to identify correct translations from wrong one. These methods are extremely language dependent.

5.1 Lexico-Syntactic Methods

This work defines the extracting translation equivalents from comparable corporawithoutrequiringexternalbilingualresources [Ote07].Tofindmeaningfulbilingualanchorswithinthecorpus, some bilingual correspondences between lexicosyntactic templates previously extracted from small paralleltexts are used. The steps involved in their method are

1. Text Processing
2. Extraction of bilingual lexico-syntactic templates fromparallel corpora
3. Extraction of word translation from comparable text using these templates

Similarity between pairs of bilingual templates is computed by taking into account their co-occurrence in each aligned segment.

$$Dice(l_1, l_2) = \frac{2 * \sum_i min(f(l_1, t_1), f(l_2, t_2))}{f(l_1) + f(l_2)}$$

They used Dice coefficient as similarity measure. Each template of the source language is linked to the most similar template of the target language provided that the Dice coefficient is higher than an empirically set threshold.

5.2 Phrasal Translation based Methods

A two-stage translation model is proposed for the acquisition of bilingual terminology from comparable corpora, disambiguation and selection of best translation alternatives according to their linguistics-based knowledge. Different re-scoring techniques are proposed and evaluated in order to select best phrasal translation alternatives [SYU03].

5.3 Dependency Relation based Extraction

The proposed approach is based on the observation that a word and its translation share similar dependency relations. It is observed that if the corpora is preprocessed with a dependency syntactic analyzer, a word in source language shares similar head and modifiers with its translation in target language, no matter whether they occur in similar

context or not [YT09]. We call this phenomenon as dependency heterogeneity. Dependency heterogeneity means a word and its translation share similar modifiers and head in comparable corpora. The modifiers and head of unrelated words are different even if they occur in similar context. Based on this observation,authors have proposed an approach to extract bilingual dictionary from comparable corpora. Not only using bag-of-words around translation candidates in context-based approach, the proposed approach utilizes the syntactic analysis of comparable corpora to recognize the meaning of translation candidates. Besides, the lexical information used in the proposed approach does not restrict in a small window, but comes from the entire sentence.

6 Enhancing the Comparability of Bilingual Corpora

The problem of enhancing the comparability of bilingual corpora in order to improve the quality of bilingual lexicons extracted from comparable corpora [LGA11] is addressed here. A clustering-based approach for enhancing corpus comparability which exploits the homogeneity feature of the corpus, and finally preserves most of the vocabulary of the original corpus. The comparability measure used extensively for comparable corpora research is introduced by this paper. This is explained in details in next section.

7 Comparability Measure

In order to measure the degree of comparability of bilingual corpora, we make use of the measure M developed in [GRM+04]

The measure proposed is based on the expectation of finding the translation for each word in the corpus. Notations used are as follows.

$$C : The\ Comparable\ Corpus$$
$$C_e, C_v : The\ English\ and\ Foreign\ language\ part\ of\ the\ corpus$$
$$C_e^V, C_f^V : The\ vocabularies\ of\ English\ and\ Foreign\ corpus.$$
$$T_w : The\ translation\ set\ of\ word\ W.$$
$$\sigma : The\ function\ indicating\ the\ presence\ of\ translation\ in\ the\ vocabulary.$$
$$\sigma(w, C^V) = \begin{cases} 1, & T_w \cap C^V \neq 0 \\ 0, & \text{otherwise} \end{cases}$$
$$M_{ef} : The\ Comparability\ Measure$$
$$M_{ef}(C_e, C_f) = E(\sigma(w, C_f^V)|w\epsilon C_e^V) = \sum_{w\epsilon C_e^V} \underbrace{\sigma(w, C_f^V).P(w\epsilon C_e^V)}_{A_w} = \frac{|C_e^V|}{|C_e^V \cap D_e^V|} \sum_{w\epsilon C_e^V \cap D_e^V} A_w$$

Where D_e^V is the English part of the given , independent bilingual dictionary.
To avoid bias towards common words we use presence/absence as a criteria rather than the number of occurrences and thus obtain.

$$M_{ef}(C_e, C_f) = \frac{1}{|C_e^V \cap D_e^V|} \sum_{w \epsilon C_e^V \cap D_e^V} \sigma(w, C_f^V)$$

The above formula shows that the metric is proportional to the number of English words translated into the foreign language part of the comparable corpora. Using a similar expression for $M_f e$ we obtain a symmetric version of the measure $M(C_e, C_f)$ obtained by considering the proportion of the words for which a translation can be found in the corpus

$$\frac{\sum_{w \epsilon C_e^V \cap D_e^V} \sigma(w, C_f^V) + \sum_{w \epsilon C_f^V \cap D_f^V} \sigma(w, C_e^V)}{|C_e^V \cap D_e^V| + \left|C_f^V \cap D_f^V\right|}$$

8 Summary

We have discussed various methods of extraction of bilingual information from comparable corpora. This bilingual information may be in form of words pairs, phrases or parallel sentences. A comparison between these methods might have been done, but it is not possible because of the diverse language pairs used by this approaches. Moreover the output of these systems are not consistent. Some of them are using precision recall metrics, where as others are using manual methods to verify their results. Few of them are also considering BLEU as the metric to judge their performance. One has to decide which approach to use based on the requirement.

References

[BD05] C. Brockett and W.B. Dolan. Support vector machines for paraphrase identification and corpus construction. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP)*, pages 1–8, 2005.

[CZ02] Y.C. Chiao and P. Zweigenbaum. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, pages 1–5. Association for Computational Linguistics, 2002.

[DGS02] H. Déjean, É. Gaussier, and F. Sadat. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

[FC04] P. Fung and P. Cheung. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In *Proceedings of EMNLP*, pages 57–63, 2004.

[FM97] P. Fung and K. McKeown. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202, 1997.

[Fun95] P. Fung. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 173–183, 1995.

[Fun98] P. Fung. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. *Machine Translation and the Information Soup*, pages 1–17, 1998.

[FY98] P. Fung and L.Y. Yee. An ir approach for translating new words from non-parallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 414–420. Association for Computational Linguistics, 1998.

[GRM+04] E. Gaussier, J.M. Renders, I. Matveeva, C. Goutte, and H. Dejean. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 526–es. Association for Computational Linguistics, 2004.

[LGA11] B. Li, E. Gaussier, and A. Aizawa. Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 473–478. Association for Computational Linguistics, 2011.

[MFM04] D.S. Munteanu, A. Fraser, and D. Marcu. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL*, pages 265–272, 2004.

[MM06] D.S. Munteanu and D. Marcu. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 81–88. Association for Computational Linguistics, 2006.

[Ote07] P.G. Otero. Learning bilingual lexicons from comparable english and spanish corpora. *Proceedings of MT Summit XI*, pages 191–198, 2007.

[Rap99] R. Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics, 1999.

[SN04] L. Shao and H.T. Ng. Mining new word translations from comparable corpora. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 618–es. Association for Computational Linguistics, 2004.

[SYU03] F. Sadat, M. Yoshikawa, and S. Uemura. Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2*, pages 141–144. Association for Computational Linguistics, 2003.

[YT09] K. Yu and J. Tsujii. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 121–124. Association for Computational Linguistics, 2009.