Real-Time Identification of Parallel Texts from Bilingual Newsfeed

David Nadeau and George Foster {@nrc-cnrc.gc.ca}

Interactive Language Technology,
National Research Council Canada,
Gatineau, Québec, Canada, K1A 0R6.

Abstract

Parallel texts are documents that present parallel translations. This paper describes a simple method that can be deployed on a real-time news feed to create an infinitely growing source of parallel texts in French and English. Our experiment was lead on the Canada Newswire news feed. Given some of its intrinsic properties, it was possible to deploy a relatively simple text matching techniques that rely on language independent cognates such numbers, capitalized words, punctuation and new lines characters. On three week of press releases, our system correctly identified the vast majority of parallel press release. It committed only minor errors on repeated news items.

Introduction

Parallel texts (also known as bitexts) are documents that present parallel translations, usually in two languages. It has been discussed (Resnik and Smith, 2003) that parallel text can be though of as a critical resource in many computational linguistic tasks like machine translation, cross language information retrieval and lexical acquisition. The problem of parallel texts availability is one of amount and freshness. This paper describes a simple method that can be deployed on a real-time news feed to create an infinitely growing and always fresh source of bitexts in French and English. It is inspired on previous works that consists in searching bilingual texts on the web. The following section presents some background works and motivation. The section 3 presents the news feed used in this research. Section 4 gives the complete algorithm to match document and create a bitexts corpus. Section 5 shows an evaluation of the method accuracy. Finally we conclude and point to areas of future research.

Background Work

To overcome the limited availability of bitexts, many researches examined the idea of automatically create parallel corpus from the web. Among these attempts, BITS is a system (Ma and Liberman, 1999) that mines the web and automatically extract bitexts.

To our knowledge, most of previous work in this area uses linguistic resources to perform an accurate match between parallel texts. In the system mentioned above, a search within all pages of a given web site is performed. This creates a lot of candidate pairs with potentially quite similar content. For this reason, techniques for extracting the valid pairs use bilingual dictionaries or statistical models to asses that translated pages share a certain amount of vocabulary. In our work, due to the intrinsic properties of the news feed under examination, we make no use of linguistic resources having, thus, a quite simple method.

Apart from web-specific preoccupations, authors of BITS note that some cue won't change in translated versions of a same text. That is numbers and some named entities. That is the basis of our method. It is also inspired from Simard *et al*. technique (Simard *et al*., 1992) that explore the intuition that translation preserves some particular punctuation like the period and the parenthesis.

A source of motivation for our work can be found in United Stated 2004 Information Technology Research for National Priorities (NSF, 2004). It is indeed explicitly explained that robust machine translation that "rapidly adapt to changes in times of crisis" is required. Even if our system was not build in response to this document, we believe that Canada's bilingual news feed has a great potential for creating timeliness corpus that are, as mentioned, crucial in machine translation research.

Canada NewsWire (CNW) News Feed

Canada NewsWire is a resource for time-critical news and information from more than 10,000 sources coast to coast and around the world. The exploited feature of this news feed is that many companies publish their communication in French and English for the bilingual audience in Canada.

Looking at three weeks of Newswire feed, it was naively noted that all press release that were publish in two languages were published in a 24 hour interval. While there's no guarantee that it hold for any possible cases past or future, it seem reasonable that the vast majority of press releases will follow this rule.

It may look like an artificial problem to listen at a news feed to identify bitexts. It could be thought that Canada Newswire may tag their releases as being translation of each other. However, Canada Newswire is only an aggregator and publisher of releases. Company sends their stories in an unordered manner, without a specified timeframe and never identifying translated versions and original ones.

Before to dig in our method for finding parallel texts in this news feed, let's note two difficulties that are challenging. These two cases can fool our system:

1) Companies can release a revised version of their press release in a short time frame, creating a conflicting 1-n bitext association

2) Company can release almost identical version of news targeted at various audiences. For instance Expeadia ever released multiple versions of news explaining that "Quebecers", "Ontarians", "British Columbian" people will take their vacation in Canada this year. All releases were almost identical except for the province.

On the other hand, an external knowledge helps us identifying the language of sources texts. Indeed, Canada newswire explicitly mention the language of all their press releases.

A Method to Continuously Build a Parallel Text Corpus

Listening to an input feed from Canada Newswire, we choose to consider every French press releases (typically less numerous than English ones). For these, the English news published 12 hours before and 12 hours after are considered as a candidate parallel text. This 24 hours time frame reduced considerably the number of candidates to look at. Also, some intuitive rules let us think that companies rarely release more than one story at a time. These favourable constraints led the following method.

Given a French text, let's consider all the English texts that were released in a 24-hours time frame:

For both documents D_i of a pair (D_1, D_2), a text model is built by creating an un-ordered bag of features.

$$TextModel(D_i) = \{m_1^i, m_2^i, ..., m_k^i\} = M_i$$

Where m_j^i is a feature of the document D_i among the following:

1. Capitalized words that are not at sentence beginning.
2. Numerals (any characters sequence starting with a digit).
3. Punctuation sings among ", (,), [,].
4. Sequences of two or more new line characters (approximation of paragraphs).

Note that a text model is built using one and only one of the features above. Four text models can therefore be built for each text.

A bag of features holds all document features along with repetitions. Given the text models, two frequency vectors $V(D_1)$ and $V(D_2)$ can be created where all features from both texts are considered.

$$V(D_i) = \left\langle f(\alpha, D_i) \middle| \alpha \in M_i \cup M_j \right\rangle$$

Where $f(\alpha, D_i)$ is the frequency of the feature α in the document D_i.

The similarity of the pair (D_1, D_2) is then computed using the cosine measure of $V(D_1)$ and $V(D_2)$ vectors:

$$sim(V(D_1), V(D_2)) = \frac{V(D_1) \otimes V(D_2)}{\|V(D_1)\| \times \|V(D_2)\|}$$

Where $\otimes$ stands for the vectorial product and $\|V(D_i)\|$ is the length of the vector D_i.

When the similarity is higher than a given threshold, the candidate pair is chosen as a bitext. The following section discusses the accuracy of bitext identification for the four features: capitalized words, numerals, punctuation and new lines.

Evaluation

Since it is out of the scope to compare the accuracy of our methodology on texts mined from the web or on other similar benchmark, we present here an evaluation of the system using individual features and a combination of them. It gives insight on what are the most useful non-changing cues in texts under the conditions that we stated.

Corpus Details: Our test corpus represents three weeks of press releases from Canada Newswire. A tenth of industries were chosen out of 65 available, ranging from technology, politics and education. The corpus is composed of 608 items (276 in French, 332 in English). We manually found 174 parallel documents (87 pairs). As explained above, these parallel texts can be found in a time frame of 24 hours. If we consider 48 hours around each French press release, 4 more parallel texts can be found. We will neglect this fact in the following experiment.

Numeral Feature: Numeral includes important features like the publication date, monetary amounts, etc. It can reach a perfect precision (100%), never matching invalid parallel texts while keeping 85% of recall. The optimal F1 (0.92) is reached at this point. Figure 1 presents the precision-recall curve when the cosine measure threshold varies:

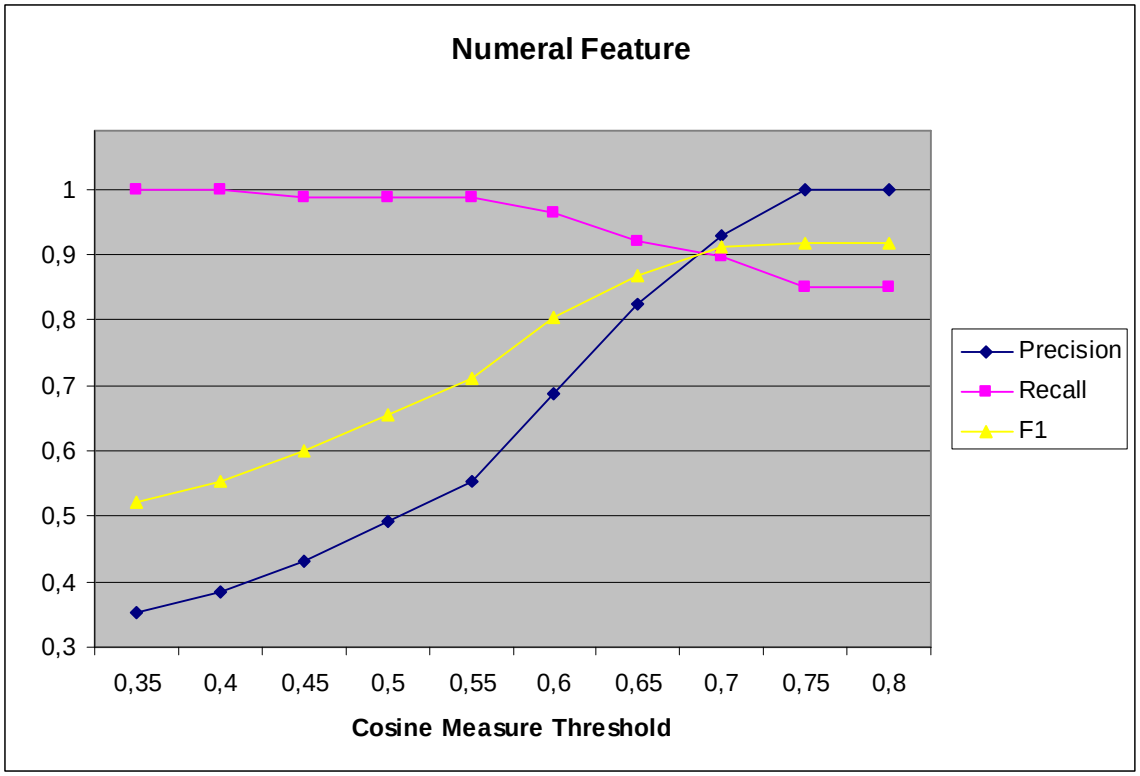


Figure 1: Numeral Feature Precision and Recall

Capitalized Words Feature: We decided to approximate Named Entities. To do so, we use capitalized words but we remove those who start sentences. This is quite an imperfect approximation of named entities but to a certain extend, it is purely statistical and language independent (for French, English, Spanish, and other language where named entities are capitalized). When two texts share the same list of capitalized word, we note a high precision in the match. However, the recall drops rapidly, indicating that many Entities are shared. This is the case since many stories are regional and share provinces and cities names. Also, we did not filter any capitalized words and the statistics are diluted by Canada Newswire standard header and trailer. Figure 2 show the capitalized words curve with a precision of 100% for a recall of 70%.

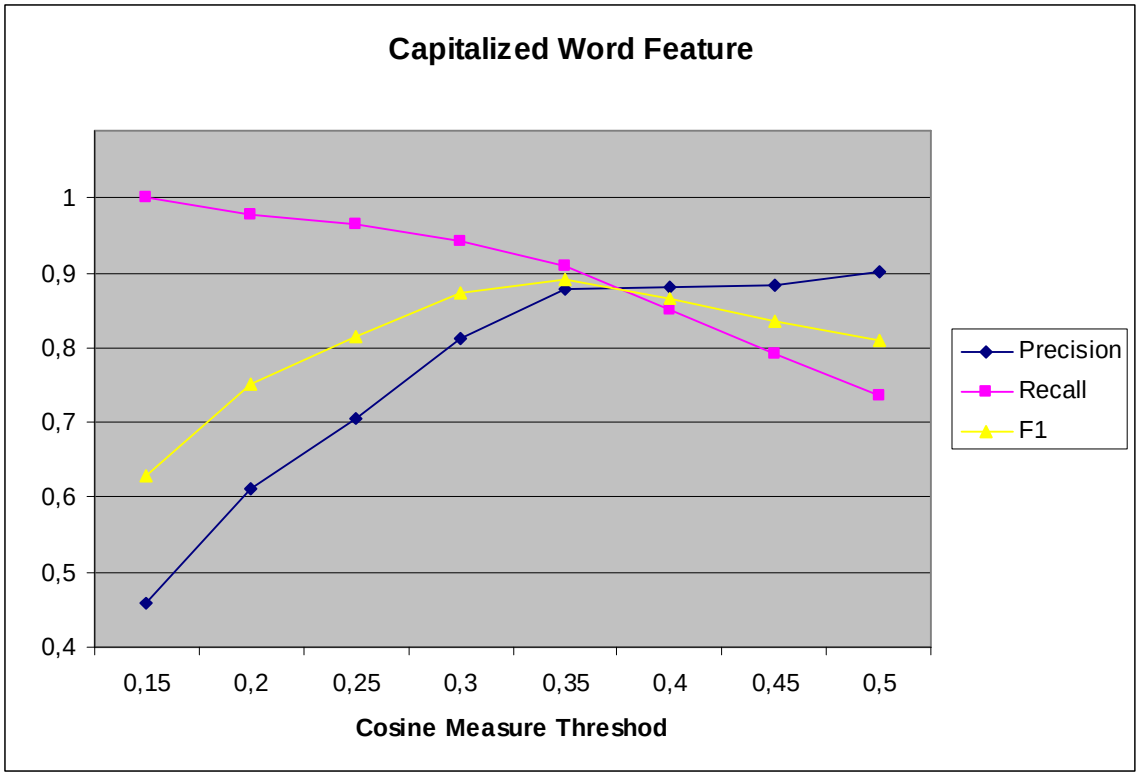


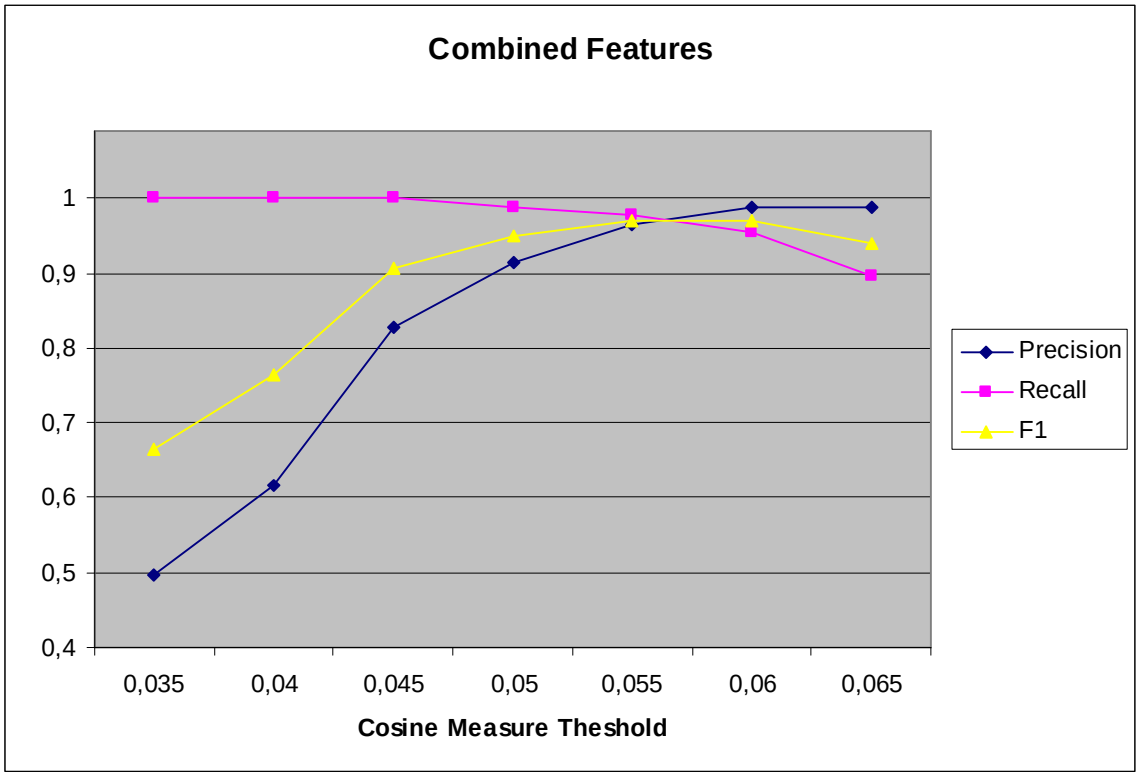
Figure 2: Capitalized Word Feature Precision and Recall

Quotes, Parenthesis and New Lines: Quote and parenthesis punctuation signs were examined because of their usage in the news. These punctuations seem to carry more invariable semantics that grammar specific comma or period. When a person is cited or a company symbol is given between quotes, this is similar in both languages. On the other side, linefeed approximates the number of paragraphs of the source text. It was believed that a text and its translation have the same length and same paragraph structure (Simard *et al.*, 1992). Those two features were mixed here because we remark a high similarity of behaviour in our experiments. It can be note that this feature do not vary a lot and therefore all texts are quite similar. Indeed, we need a high cosine measure threshold to discriminate among text pairs (0.99 and more). At this level the recall is perfect (100%), supporting the hypothesis that bitexts share punctuation and paragraphs. However, a really low precision is also noted (around 30%), meaning that too many texts shares theses features. For this reason, these attributes were rejected.

Combination of Capitalized Words and Numerals: Best results were achieved when combining capitalized words and numeral features presented above. To do so, a linear combination was applied. The similarity measures of both features were given a weight according to following this formula:

$$Sim_{combined} = a \bullet sim_{numeralfeature} + b \bullet sim_{capitalizedwordsfeature}$$

Some values for the coefficients were tested and the best we found is $a = 0.60$ and $b = 0.40$. It seems to be consistent with the result above to give a higher contribution to the numerical feature since it individually reaches a higher F1.



The optimal F1 score of 0.97 is reached when both precision and recall meet that mark. Such a result signifies that the model correctly finds 84 correct pairs out of 87 but doing so, it introduced 3 wrong pairs. Those 3 wrong guess are highly similar press releases that share almost exactly the same vocabulary but that we won't manually match. In one case the same company releases two stories about the same product but talking about different features. The matching dates, company name, product, spokesman and embedded company description with lot of numerical features lead to this error. The two other cases are repeated press releases (release submitted twice with minor correction).

Conclusion

In this paper, we presented a method for automatically build a corpus of parallel texts using the Canadian Newswire news feed. The motivation behind this work is creating a large scale corpus of bitexts for task such as machine translation or cross language information retrieval. An important aspect of our corpus is the timeliness (the property of containing the vocabulary from the latest news). Given intrinsic properties of the selected news feed, it was possible to deploy a relatively simple text matching techniques that rely on non-changing features of translated texts: numbers, capitalized words, some

punctuation and new lines characters. To a certain extend, this method can be considered as language independent. It holds as long as languages uses Arabic numbers and named entities are capitalized, in contrast with non-capitalized nouns. On three week of press releases, our system committed only minor errors on repeated news items and bulk releases by a same company with minor changes. Future works imply looking at the quality of this news feed to create parallel text at the sentence level. Bilingual applications where time is a strong constraint are also envisioned, like cross language information retrieval of news items.

References

Canada Newswire, http://www.newswire.ca/

Ma, Xiaoyi, and Liberman, Mark, Y., 1999, BITS: A Method for Bilingual Text Search over the Web, in *Machine Translation Summit VII*, September.

NSF (National Science Foundation), 2004, US Information Technology Research for National Priorities, *The National Science Foundation*.
http://www.nsf.gov/pubs/2004/nsf04012/nsf04012.htm

Resnik, Philip, and Smith, Noah, A., 2003, The Web as a Parallel Corpus, in *Computational Linguistics, Special Issue on the Web as Corpus*, 29 (3), pp. 349-380

Simard, M., Foster, G. and Isabelle, P., 1992, Using Cognates to Align Sentences in Bilingual Corpora, in *Proceedings of the Fourth International Conference on Theoretical and Methodogical Issues in Machine translation*, Montreal, pp. 67-81.