Trillions of Comparable Documents

Pascale Fung, Emmanuel Prochasson and Simon Shi

Human Language Technology Center
Hong Kong University of Science & Technology (HKUST)
Clear Water Bay, Hong Kong
{pascale,eemmanuel,eesys}@ust.hk

Abstract

We propose a novel multilingual Web crawler and sentence mining system to continuously mine and extract parallel sentences from trillions of websites, unconstrained by domain or url structures, or publication dates. The system is divided into three main modules, namely Web crawler, comparable and parallel website matching and parallel sentence extraction. Previous methods in mining parallel sentences from the Web focus on specific websites, such as newspaper agencies, or sites sharing the same URL parents. The output of these previous systems are limited in scope and static in nature. As the Web is boundless and growing, we propose to continuously crawl the Web and update the pool of parallel sentences extracted. One main objective of our work is to improve statistical machine translation systems. Another objective is to take advantage of the heterogeneous website documents to discover parallel sentences in henceforth undiscovered domains and genres, such as user generated content. We investigate a host of recall-oriented vs precision-oriented algorithms for comparable and parallel document matching, as well as parallel sentence extraction. In the future, this system can be extended to mine other monolingual or bilingual linguistic resources from the Web.

1. Introduction

As statistical approaches become the dominant paradigm in natural language processing, there is an increasing demand for data, more data, and yet more data. Just little more than a decade ago, "large corpora" used to mean a collection of user manuals, or 5 years of newspaper articles. The first statistical machine translation (SMT) system using the IBM model (Brown et al., 1990) was trained on a parallel corpus of Canadian parliamentary transcriptions in English and French - the Hansard, which amounted at the time to 117,000 sentence pairs. Fast forward to 2010, state-of-the-art SMT systems are trained on tens of millions of sentence pairs consisting of hundreds of millions of words. Much of the parallel data used to train SMT systems are manually translated by professional translators. The standard rate for such an effort is about US$0.15 per word, making good SMT systems extremely expensive to build. Organizations such as the Linguistic Data Consortium have been distributing some large corpora of translated texts for research and development at a lower cost to the user than directly commissioning translators. However, as SMT systems typically perform better on texts within the same genre as its training data, general purpose, open-domain SMT systems are only attainable if the developers of such systems have access to the world's data.

In today's world, only the most powerful search companies are privy to such information. One organization with such access - Google, the world's top search engine company, whose mission is to "organize all the world's information", has access to trillions of websites, billions of email content, videos, images, speech files, and other user generated content. As of March 2009, the (indexable) Web contains at least 25.21 billion pages (World Wide Web Size, 2009). Google search had discovered one trillion unique URLs. And its translation system is statistically trained from all the data that is within its grasp. Google, while having this access, does not distribute the result of its mining to the public, except through its services. Yet, as the Web founder Tim Berners-Lee famously put it, "*The power of the Web is in its universality. Access by everyone regardless of disability is an essential aspect.*"

In this paper, we address the "disability" of statistical natural language research in general, and SMT systems in particular, to access the information on the Web as a training corpus, and propose a multilingual Web crawling and mining system as a tool to facilitate our community to mine the Web for more linguistic resources.

The World Wide Web is a "*boundless world of information interconnected by hypertext links*". We argue that the Web is a virtually infinite and continuously growing corpus for natural language processing. Rather than taking a snapshot of it at one moment, and use the result as a static corpus, we propose to continuously crawl the Web for new, comparable data for mining parallel sentences. Rather than focusing on a single domain such as news, or on translated parallel sites with matching structures, we propose to look for sites that are comparable in content, HTML structure, link structure, URL as well as in temporal distance as they potentially contain parallel sentences.

Much effort has been made in the past to try to automatically extract parallel resources from comparable corpora on one hand, and to use the Web as a corpus on the other. Both approaches (often combined) allow more diversity in the data harvested. (Resnik and Smith, 2003) directly extracted parallel texts from the Web, relying mostly on URL names. Some work has been done to extract parallel resource (sentences, sub-sentential fragments, lexicon) from comparable data. (Munteanu and Marcu, 2005) showed they can extract relevant parallel sentences using a supervised approach on newspaper corpora, although their main goal was to show how they manage to use such resources to improve Statistical Machine Translation. (Fung and Cheung, 2004; Wu and Fung, 2005) extracted parallel sentence from quasi-comparable corpora, that is corpora containing

Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010, pages 26–34
Malta, 22 May 2010

documents from the same domains as well as documents of different domains.
We need to be able to combine advanced IR/Web crawling techniques with advanced NLP methods in order to obtain large and high quality sets of parallel sentences. From this point of view, we do not want to focus on one particular domain (such as newspaper, as it is often the case in related works). Of course, we are aware and will keep in mind than better results can be obtained from certain kind of documents (for example, Wikipedia constitutes a large source of very comparable, easy to harvest and well structured documents), but propose a general approach for mining from any website, in any dominant Web language. We strive to reduce the language dependency and domain dependency to a minimum.
This is work in progress and this paper is intended as a position paper to present our objectives and arguments to the community of NLP researchers. In the next section, we take a look at the challenges that we encounter and how we plan to solve them, step by step. Section 3 describes the experimental setup and preliminary results of our experiments. We then conclude in Section 4 and discuss future directions in Section 5.

2. Challenges

Existing tools (Munteanu and Marcu, 2006; Resnik and Smith, 2003; Ma and Liberman, 1999) mine parallel sentences from a pre-defined set of archival data, with temporal and domain constraints. Some of these tools do not crawl the Web but rather, they try to mine parallel texts (Resnik and Smith, 2003) or parallel sentences (Munteanu and Marcu, 2006) from a pre-existing archive. (Ma and Liberman, 1999; Chen and Nie, 2000) developed tools that dynamically mine parallel sentences from a subset of the Web. However, these tools have become obsolete over time and the Web has since grown tremendously in the last decade. Most other methods of mining parallel sentences from comparable or parallel corpora require training from existing parallel corpora and therefore, are often only applicable to a single domain or genre. Many issues related to the challenge of mining parallel sentences from the Web has been studied and some interesting achievements have been made.
Two strategies can be adopted when mining parallel sentences: favoring recall or precision. Favoring recall will provide many pairs of sentence, but the quality of those pairs (the parallelness) is likely to be low. However parallel sub-sentential fragments (Munteanu and Marcu, 2006) can still be of great value, especially if they can be post-processed to filter out the non-parallel segments (Abdul-Rauf and Schwenk, 2009). On the other hand, favoring precision yields high quality parallel sentences (moreover, reliable alignment of sentences) at the cost of probably missing many valuable information. We focus on both approaches. For the purpose of improving statistical machine translation systems, we need to mine parallel sentences with high precision, measurable by SMT performance, not just human judgment. At the mean time, as "more data is better data" for statistical MT systems, we will also strive to improve the recall rate, while maintaining precision. We are also interested in obtaining large amounts of data quickly.
Last but not the least, even though our current objective is to mine parallel sentences from the Web, it is potentially useful to crawl the Web for other language resources, such as translation lexicons, or monolingual resources. Since the Web crawling and indexing task is non-trivial and time consuming, we need to design the system so that useful information are retained for future processing, without having to recrawl the Web for the same pages.
To summarize, we need to meet the following challenges for our task of mining parallel sentences from the Web:

1. Recall - include as many websites as possible that might contain parallel sentences
2. Precision - to be able to find high quality parallel sentences that can improve SMT performance
3. Domain and topic - to be able to find parallel sentences in as many domains/topics as possible
4. Language - to be able to find parallel sentences in different language pairs
5. Heterogenous - the system must find websites that are not just translations of each other but also others that have similar content
6. Up-to-date and always available - the system needs to crawl the Web continuously for new additional document resources
7. Query-driven - the system can accept queries to crawl and search for specific websites
8. Scalability - the system needs to be scalable to run on multiple nodes of servers in parallel.
9. Speed - fast algorithms are needed to enable us to crawl the Web efficiently for the mining task.
10. Extendable - the system needs to be modular and extendable to other mining tasks, in addition to parallel sentence mining.

The whole process is described in figure 1 and the different modules are described in the following sections.

2.1. Crawling the Web

A Web crawler is a program that automatically downloads pages from the Web. To mine parallel sentences from the entire World Wide Web continuously and automatically, a main component of our tool is a Web crawler that collects as many documents from the Web in a given language pair continuously and indexes each page for comparable document searching. The Web crawler indexes Web pages on the Web to enable them to be searchable. The main function of our system currently is to act as an comparable document search engine which discovers articles in another language that are comparable or parallel to any input text. So in the first stage, we need to crawl and index both the English Web (i.e. all English websites) and the Chinese Web. We build an index including all English pages like a search engine. When the index has reach a certain size, say 1M pages, we

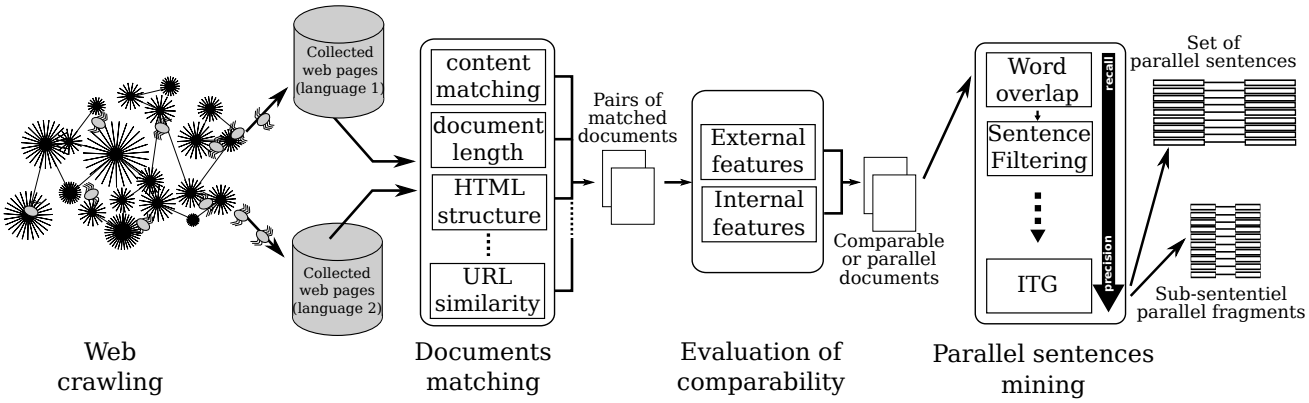

Figure 1: Overview of the sentence extraction system

will process each Chinese page to find its comparable English page in the index.

Queries to conventional search engines normally contain one or more distinct keywords. However, the query to our system at this stage is a document which may contain hundred of words. The tool searches the index and finds documents in another language that are comparable to the input. It is a high dimensional search problem with time complexity of $O(n \times m)$ where n and m are number of websites in the two languages (i.e. English and Chinese respectively). (Gionis et al., 1999) introduced a hashing method for high dimensional similarity search which can be used to reduce computation time. For our purposes, we suggest that some kind of topic or genre clustering can be carried out first to reduce the search dimension. Methods for topic classification, taking into consideration content and other information, can be used to speed up the search as well.

After we indexed a significant amount of Web pages, say 1 million pages, we start to use the search engine to get comparable documents. For each Chinese document, we first translate it into English by an MT system, such as Google Translate, or simply convert it to the word index in a bilingual lexicon. Then by searching the index we can obtain a ranked list of English texts, in terms of comparability. Those document pairs are returned as the output of the search engine. We assume that for each Chinese documents, there will be some comparable documents in English.

Simple bag-of-words comparison cannot tell us whether two pages are actually comparable, noisy parallel or parallel. So we will need other measures, described in the next section, to achieve our mining objective. In consideration of such measures, we must first index the websites accordingly. In our system, the following features are considered in the indexing step:

- Page content in terms of words
- Position of words in the document
- URL structure
- HTML structure
- Link structure
- Image file names
- Time of creation if relevant

During indexing, unlike conventional Web crawlers, we must convert all information above into index numbers. Word IDs, for example, must correspond to those in a bilingual lexicon for our source and target languages. Multiple translations of the same word can be considered. Word features such as tf/idf, frequency rank within the same page, word positions, etc. should be indexed.

In addition, the Web crawler is configured to collect different types of documents by various regular update intervals. A stochastic model for crawl target selection (Akamine et al., 2009) is implemented to control the revisit time of the crawler in order to keep the document up-to-date. For news websites, Web pages can be collected daily by the crawler while the visit frequency of other websites can be much longer.

Previously, (Chen and Nie, 2000; Yang and Li, 2004; Gleim et al., 2006) developed a parallel text mining system on bilingual websites sharing the same root URL. (Munteanu and Marcu, 2005) focused on some news websites only. They tried to extract parallel sentences from given sets of known websites without crawling the Web. Whereas the result of such work has shown to improve SMT performance, many parallel sentences exist on other websites and the sentence pairs reside on different hosts are never discovered by their more limited and static approach. (Chen and Nie, 2000) developed a tool PTMiner which mines parallel sentences under the same hostname. The Web crawler of PTMiner performs breadth first search on the same host only. In our case, we must crawl and index boundless number of websites (hostnames) continuously, rather than search for and download a part of the Web only like these previous work.

The Web crawling speed is mainly constrained by connection bandwidth. In the initial testing, we crawl the Web using 10 spiders over Ethernet, reaching the speed of one page per second. For indexing each page, a single PC with Core Duo processor at 2.0GHz is able to index 50 pages per minute. With very limited optimization, a PC running as the database server takes 10 seconds to process each Chinese document when there are 10,000 pages in the database.

We use MySQL as the central database server which is scalable to run on clusters. The Web crawlers work independently. It is possible to have several groups of spiders to crawl the Web and index pages.
We also use a black list to avoid crawling sites containing mostly non-textual material, such as YouTube, Picasa, Flicker, etc.

2.2. Matching comparable and parallel documents

To improve the recall of mining parallel sentences, we need to be able to measure and classify document pairs into not comparable, quasi-comparable, comparable, noisy parallel and parallel in order to match them better. As mentioned above, using quantitative measures, we will select documents that are comparable and noisy parallel (including parallel). According to (Fung and Cheung, 2004), quasi-comparable and comparable documents are those that were written independently but on more or less the same topic. In such cases, structural features are not useful. Noisy-parallel documents refers to a pair of source and translated document, that were either adapted or evolved in different ways. For example, Wikipedia article that was once the translation of another Wikipedia page, but evolved in time due to different contributors can be either noisy parallel or comparable to the source article.
In order to improve recall of parallel sentences between two texts, it is important to select very comparable documents but not be restricted to translated, parallel documents only. The notion of comparability is hazy and is still an open question. Practically, it depends on the expected usage of the documents. The comparability is generally evaluated on both internal and external criterion. External criterion are qualitative features, such as the topic, the domain, the time of publishing or the discourse, whereas internal criterion are quantitative features, such as the quantity of common vocabulary.
(Kilgarriff, 2001) tried to answer a related question by measuring the similarity of two corpora. He observed that such a measure is not trivial since corpora are complex and multidimensional objects. Two corpora can be close for one dimension and distant for another. In this context, the notion of similarity is connected to the notion of homogeneity in one corpus. A homogeneous corpus contains the same kind of document (Biber, 1989), that is, where some particular linguistic distinctiveness can be found. We focus on comparable documents rather than a collection of corpora. The question of homogeneity is in our case not really relevant. We therefore focus on different features, external and internal. (Fung and Lo, 1998; Fung and Cheung, 2004; Carpuat et al., 2006) previously proposed to compare the frequency rank of seed words in documents to be matched. Similar documents should have a similar representation of the common vocabulary. Such comparison can be visually evaluated, see Figure 2. Identical documents should rise a perfect diagonal, unrelated documents should show no such tendency. To quantify the similarity of documents, we also use a regression score which evaluate the dispersion of the data from the diagonal.
This score works well for documents containing a significant number of content words, but is brittle on smaller documents. If few seed words are found between two documents, the dispersion will be small, whereas documents with many common seed words might be seen more similar, since more dots will be compared. Therefore, we need to weight the raw score to get more significant information. An example is given in Figure 2: less than 50 words are common to both texts, which is too sparse for our measure. We then need to rely on other features to evaluate comparability or to be more precise, to evaluate whether two documents might contain translated sentences.

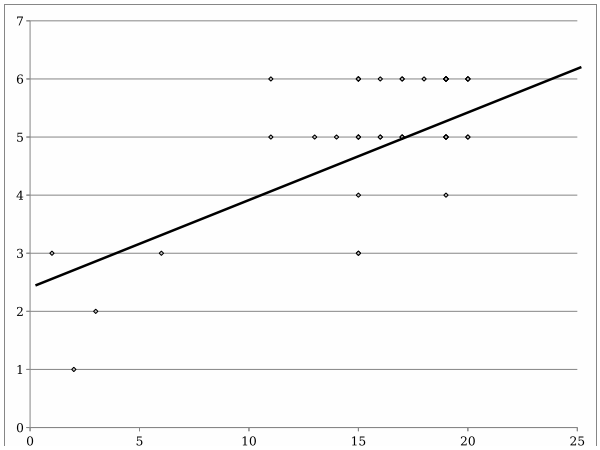

Figure 2: R^2 computation on two parallel documents about Lamma Island.

(Resnik and Smith, 2003) looked for pairs of document in translation by searching for specific link in a parent page (with links to several version of one document, in many languages) or in sibling pages (with link such as "this document in English"). We suggest that external features can be used, such as URL structure, document length, html structure, link structure, or image file names.

2.3. Mining parallel sentences

Mining parallel resources from comparable corpora has been done in several studies. (Munteanu and Marcu, 2005) proposed an approach to mine parallel sentences from selected comparable documents using a supervised Maximum Entropy classifier. One goal of their work was to rely on large amount of out-of-domain parallel data and small amount of in-domain parallel data to complete in-domain knowledge for MT. The initial parallel data are used to train the EM classifier, which will determine which sentences are good translation candidates (based on many features, starting with word overlap and length ratio of pairs of sentences). They work on newspaper data in English, Chinese and Arabic. (Fung and Cheung, 2004) looked for parallel sentences and bilingual lexicon from very non-parallel corpora, defined as collection of document on the same topic (in-topic) or not (off-topic). Rather than relying on the "*find-topic-extract-sentence*" principle (e.g. find in-domain documents, then look for translations), they proposed to "*find-one-get-more*". In other words, if parallel sentences have been found between two documents, they are likely to share more parallel sentences. They used a cosine sim-

ilarity measure to compare pairs of sentence and raised pairs above a given threshold for English/Chinese alignment. This approach raised interesting parallel resources, but they were shown to be quite scarce among unrelated documents. Furthermore, this approach applies on large amount of data.

For texts that are translations but contains a lot of noise, such as one-to-many translations, or inserted examples and graphs, or even occasional segments that are not translations of each other, we propose to adapt the DK-vec algorithm (Fung, 1998; Fung and McKeown, 1997; Fung, 1995) which use an iterative Dynamic Time Warping method to match a bilingual lexicon, used later as anchor points to align sentences. This method is interesting for it is totally unsupervised and language independant: the bilingual resources can be boostrapped from the document. Furthermore, this approach has been shown to be efficient for document without strict sentence boundary information. It was designed for noisy-parallel corpora, basically yielding a path of lexicon alignment that is not necessarily the diagonal if there is noise. DK-vec is also unique in that it uses the position feature and the (sentence) length feature implicitly in the dual objective of alignment and bilingual lexicon extraction. Other methods either use an existing lexicon and position feature to perform alignment, or use the length feature for alignment.

Finally, the results provided by high-recall method can be filtered, for example using Inversion Transduction Grammar (Wu and Fung, 2005). When using word overlap methods (or cosine similarity), sentences that share a common vocabulary but do not have the same meaning are likely to yield a high score. As an example, this pair of sentence, extracted from French newspaper *Le Figaro* and English *New York Times* obtain a high score when using word overlap:

> **En:** "National Highway Traffic Safety Administration has received about 100 complaints involving the brakes of the Prius new model."
>
> **Fr:** « Aux Etats-Unis, une centaine de plaintes ont été déposées auprès de l' administration de sécurité routière américaine pour des difficultés de freinage avec la Prius. »
>
> **Trans:** *"In the United States, about one hundred complaints have been submitted to the american administration of traffic safety for difficulties when braking with the Prius"*

Even though both sentences have roughly the same meaning, they cannot be considered parallel. ITG can then be used to take a closer look at the sentence constituent structures (predicate argument dependencies) and will eventually allow us to filter out this candidate pair, to only keep strictly parallel candidates. ITG has been shown to be efficient for this particular task and are language independant.

All in all, the overall process, from crawling the Web to parallel sentence extraction can be seen as refining a raw material (the Web) to obtain golden resources, each of the step attempting to filter out irrelevant data.

3. Preliminary Experiments

We ran an experiment to roughly evaluate the feasibility of our task by trying to extract parallel sentences from a subset of French and English Wikipedia. It is hard to precisely estimate the amount of parallel sentences available from the Web, for several reasons:

- the availability and density of parallel sentences is highly related to the type of document processed; the Web is a heterogeneous resource. It is not possible to infer an accurate estimation from a small subset evaluation.
- assuming we already had a high-recall and -precision tool to mine parallel sentence from the Web, we can not ensure we have found them all (recall estimation is, in that case, impossible). We can estimate the precision on a small subset, but the precision is also related to recall.

It would be presumptuous therefore to claim anything regarding the density of parallel sentences from the Web, however we might still want to have a look, at least to confirm that there are some, and that they can be extracted automatically.

3.1. Experimental setup

We randomly extracted 1,000 pairs of articles from French and English Wikipedia by considering articles with the exact same title (at the time we write this paper, there were 548,900 pairs of articles available). Most of these articles refer to proper names (e.g. biography of a famous figure, book titles, other works) and few of them are animal species. No distinction was made for articles that are translations or just comparable. We tried to mine parallel sentences in pairs of documents only, using a simple word-overlap measure and a French-English bilingual dictionary. The word-overlap score is evaluated based on the number of common words between two sentences, penalized by the number of words whose translation is in the dictionary and that can not be found in the other sentence. The word-overlap score is detailed in equation 1.

$$wo(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2| + |S_1 - S_2| + |S_2 - S_1|} \quad (1)$$

In equation 1, intersections/disjunctions of set is computed only on known elements, no penalty is imposed on unknown words. The sentences are cleaned to filter functional words using a list of stop words in English and French. We used a threshold to keep interesting candidates (> 0.2). This threshold is arbitrary and can be increased to maximize precision, but will allow us to observe the translation candidates.

3.2. Results

Using this experimental setup, we extracted 1,233 candidate translations. The top-ranked ones happen to be correct but are mostly useless, as they concerns short titles or structure information (typically, we obtained 29 occurrences of the correct translation *Reference*/*Références*, and

37 occurrences of *See Also*/*Voir Aussi*). We also obtained many alignment of dates or proper nouns, or alignment of irrelevant data. Moreover, due to the type of documents found in Wikipedia (especially given the constraints of selection we used), we found many "identical matches", such as *Ravat-Malvern Star*/*Ravat-Malvern Star*. This kind of alignment accounts for more than 85% of the sentences extracted. This latter observation shows that cleaning documents from the web is an issue that should not be underestimated. We need to ensure to process *content* of webpages, and make sure to get rid of useless information such as menus or advertising.
Apart from short sentence alignment, we can classify other candidates into three groups:

- Exact parallel sentences. Same meaning, same organisation of the sentence, same amount of information.
- Partially parallel sentences. One sentence is likely to contain more information, or they are organised differently. Those can still be of interest if they can be post-processed.
- False Positives. Sentences that were matched but don't share a common meaning.

Surprisingly, we obtain very few false positives with a score higher than 0.25. One example is given below:

> **English:** The DC2 Type R was the only Type R ever sold in North America (With the Acura badge)
>
> **French:** Les Honda Type R sont les modèles sportifs les plus performants du constructeur Honda automobile.
>
> **Translation:** *The Type R Honda are the most performant race models from the Honda motor company.*

Overall, we found about 12 true wrongly aligned sentences only. This is a very interesting result, since it shows that, as long as we ensured that i) the documents we are processing are strongly comparable and ii) we found some translation candidates with a high enough score, those candidates are reasonably reliable. We found about 150 parallel or partially parallel sentences. They were manually classified and some examples are given in Tables 1 and 2.
These results are interesting because they show a reasonable amount of parallel sentences can be found. However, as we emphasized previously, these results can not help us evaluate precision/recall or ratios of parallel sentences among documents from the Web. Our ultimate goal is not to harvest parallel sentences from Wikipedia, in French and English. Some effort will be necessary to obtain more interesting results from the rest of the Web.

4. Conclusion

We argue that it is possible to mine a heterogenous corpus of parallel sentences in the dominant Web languages, in any domain and any topic, from the Web. We propose to combine sophisticated information retrieval methods with statistical natural language processing methods to better harvest the material from the Web. Many assumptions made by previous work do not hold as we move from mining from limited domain, and limited genre websites to the entire Web. We suggest that an optimal combination of recall-oriented algorithms and precision-oriented ones will enable us to mine the gold nuggets - linguistic resources - in the information ocean that is the World Wide Web. The Web is boundless and amorphous. The innovation of our proposed work lies in our consideration of the Web as a dynamic, time-variant corpus, rather than a static archive. We propose a combination of content, structural, and temporal features to crawl the Web with the objective of continuously mining useful multilingual linguistic resources such as comparable or parallel corpus. We suggest to investigate a host of recall vs precision-oriented methods to mine parallel sentences from comparable websites returned by our Web crawler. Some initial experimental results have been shown as the existence proof of parallel sentence pairs in non-parallel websites, such as the Wikipedia.

5. Discussion, future work

This project is large and ambitious, and each step will require extensive study of state-of-the art approaches, and hopefully improvement of previous approaches. As we mentioned, many of the assumptions made previously might or might not hold for a project at this scale. For example, relevant documents that are useful for cross-lingual retrieval, based on page-ranked search results, might not contain any parallel sentences. Websites that are not translations of each other, might still contain parallel segments. Extraction systems using classifiers and rankers trained from an in-domain corpus are not applicable to our system as we do not focus on any specific domains. Nevertheless, it can be useful to classify the final extracted sentences into different domains for training domain-specific SMT systems.
With the rising popularity of Web 2.0 and Web 3.0 websites, there are more and more user generated content on the Web and many of them relate to each other in very interesting ways, such as user feedback on the latest Apple products, fan club discussion on the latest gossip of a celebrity. Such topics are temporal in nature - and available in multiple languages. Our system downloads and compares these websites as part of its output. We would like to analyze the results and see whether such data can be used to improve an SMT system on user generated content.
The Semantic Web is another effort by the W3C community to improve upon the current HTML annotation of Web pages to include the "meaning" of Web content for Web browsers and search engines to better "understand" and satisfy user queries. When mature, the new semantic annotation scheme can potentially provide a new feature, the semantic feature, to our system in mining and comparing websites.
A problem that remains to be addressed by our system is that there are many more parallel (and other) data available on the Web than those indexed by a search engine or by our system - there are compressed files of translated texts, such as the United Nations Parallel Corpus, or image files of scanned documents, such as books in translated into multiple languages, contents of tables, subscription-based

French	English
Histone H4, un composant de la structure de plus haut niveau de l'ADN des cellules eucaryotes	Histone H4, a component of DNA higher structure in eukaryotic cells
L'important engagement d'Henry Ford à réduire les coûts aboutit à de nombreuses innovations techniques et commerciales, notamment un système de franchise qui installe une concession dans toutes les villes en Amérique du Nord et dans les grandes villes, sur les six continents.	Henry Ford's intense commitment to lowering costs resulted in many technical and business innovations, including a franchise system that put a dealership in every city in North America, and in major cities on six continents.
Dans celui-ci les angles sont confinés à un plan ; donc l'étape suivante devrait être une algèbre quadruple quand l'axe du plan devient variable.	In it the angles are confined to one plane ; hence the next stage will be a quadruple algebra, when the axis of the plane is made variable.
Le segment six a un motif semblable mais avec moins de bleu et le segment sept est presque entièrement noir, avec seulement une fine bande bleue à la base.	Segment six has a similar pattern but with more restricted blue and a broader area of black, and segment seven is mostly black, with just a narrow blue area at the base.
Swami Shivananda Saraswati (8 septembre 1887 - 14 juillet 1963) est un maître spirituel hindou très réputé et un promoteur du Yoga et du Vedanta.	Swami Sivananda Saraswati (September 8, 1887 July 14, 1963) was a Hindu spiritual teacher and a well known proponent of Sivananda Yoga and Vedanta.

Table 1: Sample of parallel sentences extracted.

French	English
L'album est sorti le 18 novembre 2009 sous le label Regain Records.	**The album was** officially **released on November 18, 2009 via Regain Records.**
De 1977 à 1981, il travaille dans l'équipe la Commission des vétérans à la Chambre des représentants.	**From 1977 to 1981,** Webb **worked on the staff of the House Committee on Veterans Affairs.**
Elle **donne** à toute personne recevant le logiciel **le droit** illimité **de l'utiliser, le copier, le modifier, le fusionner, le publier, le distribuer, le vendre et de changer sa licence.**	The MIT License **states** more explicitly **the rights** given to the end-user, including the right **to use, copy, modify, merge, publish, distribute, sublicense, and/or sell the software.**

Table 2: Sample of partially parallel sentences extracted.

websites etc. This Deep Web (or Hidden Web) is orders of magnitudes larger than the visible Web. The current Web reachable by search engines is about 167 terabytes whereas the Deep Web is estimated to be 91,000 terabytes. Whereas developing a comprehensive tool to crawl the Deep Web is perhaps beyond the scope of our proposed work, for a specific natural language task, such as SMT, we might want to dig deeper into a specific genre of data.

One of the most interesting part, and a cornerstone of this work is the ability to evaluate comparability. This is a particularly tricky question, since the comparability concept itself is hazy. Some assume than noisy-parallel corpora are comparable, some assume that document in each languages has to be written independently while others claim there is a continuum from non-related to parallel corpora. Quantitatively and qualitatively evaluating the comparability might bring to light a more precise definition of comparability and comparable corpora. For websites, structural comparability does not necessarily lead to content comparability. Given the large amount of websites, should we first constrain our search with URL structural matching as in (Resnik and Smith, 2003)? Or should we start with the least stringent criteria for recall? We argue for the latter. All Wikipedia articles have similar URL names and HTML structures, but with very different content. For example, Chinese Wikipedia is clearly not a translation of the English Wikipedia.

As we mentioned that our system aims to help users mine multilingual resources from the Web for more than one applications. As an example, one of the main interest in comparable corpora concerns bilingual lexicon extraction, which is generally performed on large corpora (millions words (Fung, 1995; Rapp, 1995)) following *the more data is better data* principle, or relying on smaller but more constrained, specialized corpora (Daille and Morin, 2005; Chiao and Zweigenbaum, 2002) to focus on terminology. Both approaches fail to find relevant translations for rare words, for two reasons: (1) Even in large corpora, there is no guarantee that a source word will occur in the target corpus (Zip's law); (2) these approaches mostly rely on context-based comparison - a word and its translation are likely to have similar contexts, just as a word and its synonyms share the same context (following the Firthian principle that "you shall know a word by the company it keeps" (Firth, 1957)). Rare words by definition do not occur frequently enough to create a meaningful context and cannot be compared efficiently. (Pekar et al., 2006) tried to circumvent this issue by smoothing the context of rare words using the context of their k-nearest neighbors. They obtained a significant improvement in the quality of the lexicon alignment, by lowering the rank of correct translation candidates.

This raises another interesting question: Are there rare words in the Web? Does the notion of hapaxes still exist? There is a direct answer: yes, of course. First, one can invent a word that could not be found anywhere else, but this is a trivial case. Rare words occur in languages that are scarcely represented on the Web. Apart from these cases, can we assume than all the words and terms of the world's top Web languages can be found on the Web? A related experiment done by our group found that all Chinese named entities in the Wikipedia pages are translated into English somewhere on Chinese websites. A simple regular expression search can return the translation results.

Finally, rather than relying on large quantities or highly constrained corpora, we believe we can take advantage of the diversity and availability of comparable documents (and typically, take advantage of the availability of comparable documents in many languages, to perform multi-source alignment). A lexicon acquired in such a way can be used as feedback to the whole sentence alignment process, to increase the quality of word overlap estimation and comparability evaluation, raising better matched documents and higher quality parallel sentences.

6. References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 16–23.

Susumu Akamine, Yoshikiyo Kato, Daisuke Kawahara, Keiji Shinzato, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara. 2009. Development of a large-scale web crawler and search engine infrastructure. In *Proceedings of the 3rd international Universal Communication Symposium (IUCS'09)*, pages 126–131.

Douglas Biber. 1989. A typology of english texts. *Linguistics*, 27:3–43.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistic*, 16(2):79–85.

Marine Carpuat, Pascale Fung, and Grace Ngai. 2006. Aligning word senses using bilingual corpora. *ACM Transactions on Asian Language and Information Processing*, 5(2):89–120.

Jiang Chen and Jian-Yun Nie. 2000. Parallel web text mining for cross-language information retrieval. In *Recherche d'Informations Assistée par Ordinateur (RIAO)*, pages 62–77.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212.

Béatrice Daille and Emmanuel Morin. 2005. French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, pages 707–718.

John Firth. 1957. *A synopsis of linguistic theory 1930-1955*. Studies in Linguistic Analysis, Philological. Longman.

Pascale Fung and Percy Cheung. 2004. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In Dekang Lin and Dekai Wu, editors, *Proceedings of Empirical Methods on Natural Language Processing (EMNLP'04)*, pages 57–63, Barcelona, Spain.

Pascale Fung and Yuen Yee Lo. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of COLING-ACL98*, pages 414–420.

Pascale Fung and Kathleen McKeown. 1997. A technical word- and term-translation aid using noisy parallel corpora across language groups. *Machine Translation*, 12(1/2):53–87.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In David Yarovsky and Kenneth Church, editors, *Proceedings of the 3rd Workshop on Very Large Corpora (VLC'95)*, pages 173–183.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16.

Aristides Gionis, Piotr Indyk, and Rajeev Motwani. 1999. Similarity search in high dimensions via hashing. In *VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases*, pages 518–529, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Rüdiger Gleim, Alexander Mehler, and Matthias Dehmer. 2006. Web corpus mining by instance of wikipedia. In *WAC '06: Proceedings of the 2nd International Workshop on Web as Corpus*, pages 67–74, Morristown, NJ, USA. Association for Computational Linguistics.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.

Xiaoyi Ma and Mark Liberman. 1999. Bits: A method for bilingual text search over the web. In *Proceedings of Machine Translation Summit VII*, page 6.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallelcorpora. In *21st International Conference on Computational Linguistics (ACL'05)*.

Viktor Pekar, Ruslan Mitkov, Dimitar Blagoev, and Andrea Mulloni. 2006. Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL'95)*, pages 320–322.

Philip Resnik and Noah A. Smith. 2003. The Web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

World Wide Web Size. 2009. `http://www.worldwidewebsize.com/`.

Dekai Wu and Pascale Fung. 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *IJCNLP*, pages 257–268.

Christopher C. Yang and Kar Wing Li. 2004. Building parallel corpora by automatic title alignment using length-based and text-based approaches. *Information Processing & Management*, 40(6):939 – 955.