Correlation between the quality of focused crawlers and the linguistic resources obtained from them

Abstract. *Focused web crawlers have been used for the automatic acquisition of lexical resources for particular domains, gathering websites related to a set of topics of interest. For this purpose, a portion of the web graph is traversed, and the documents corresponding to pages considered relevant are stored and treated as a corpus. It is important to traverse this graph in a targeted way, organizing pages in a queue that prioritizes pages that are more likely to be relevant. Texts collected by these tools can be used to train domain-specific machine translation (MT) systems. In this work, we compare the performance of focused crawling algorithms, measured with standard metrics, and the quality of the linguistic resources obtained, in order to try to establish a correlation between both. Also, we propose a novel, extrinsic metric to evaluate the efficiency of a focused crawling algorithm.*

Palavras-chave: focused crawling, lexical resources, machine translation, correlation

Aluno: Bruno Rezende Laranjeira (bruno.rezendelaranjeira@inf.ufrgs.br)

Orientadora: Viviane Pereira Moreira (viviane@inf.ufrgs.br)

Co-orientadora: Aline Villavicencio (avillavicencio@inf.ufrgs.br)

Nível: Mestrado

Programa de Pós-Graduação em Computação

Instituto de Informática - Universidade Federal do Rio Grande do Sul

Ingresso em: Primeiro semestre de 2012

Época esperada de conclusão: Segundo semestre de 2013

Etapas concluídas: Foi construída e testada uma ferramenta extensível para coleta de documentos na web. Alguns algoritmos de coleta focada também já foram implementados. Procedimentos para a continuação dos experimentos já estão definidos

Etapas futuras: Finalizar a implementação de outros algoritmos de coleta focada. Treinamento do sistema de tradução automática. Avaliação das traduções geradas. Análise da correlação entre as duas medidas.

1. Introduction

Web crawlers [Liu 2009] are used by web search engines, like Google or Yahoo, to collect documents and use them to construct their indexes. To do this, they traverse the web graph, fetching pages and storing the information necessary for the intended task (e.g. text, hyperlinks, figures, etc.). Crawlers start collecting an initial set of seed pages and keep all the outlinks contained in them in a queue known as *URL frontier*. This process continues recursively, visiting pages contained in the URL frontier, whose links are extracted and added to the frontier, until the frontier is empty or a stopping criteria, which may be a given number of pages to be collected or a maximum time allowed for the crawl, has been reached. Besides their use on web search engines, crawlers can also be used in any other computational problems that need gathering web pages, like question answering systems and for using the web as a corpus.

Depending on their purpose, surface web crawlers can be classified in two categories [Liu 2009]: *universal crawlers*, which aim to collect all kinds of pages, and *focused crawlers* which intend to gather only pages belonging to a set of topics of interest. Focused web crawlers reveal themselves very useful for acquiring corpora about a given domain. Corpora gathered by a focused crawler do not have to be limited to a single language. A *multilingual comparable corpus* can be used to train topic specific machine translation (MT) systems. In this work, we treat *comparable corpora* as sets of texts about a given subject written in more than one language. They differ from parallel corpora in that they do not need to be exact translations of one another.

In this work, we analyze the correlation between the performance of a focused crawling algorithm and the quality of the linguistic resources that can be extracted from the gathered corpora. To estimate the quality of the resources, we use a focused crawling algorithm to collect comparable corpora and use them to train a domain-specific MT system, and then evaluate the generated translations, which must be about the same domain that guided the crawl. We also propose to use the quality of the lexical resources obtained as a novel measure to evaluate focused crawling algorithms. Since we believe that the quality of the translations generated by the trained MT system reflects the quality of the lexical resources obtained, this seems to be a reliable, although extrinsic, metric to assess the performance of focused crawling algorithms.

The remainder of this paper is organized as follows: In Section 2, we introduce the background knowledge necessary to understand our methodology and experiments. Section 3 presents related work, important to the development of this one. Sections 4 and 5 bring, respectively, the details of our methodology and of our preliminary experiments. Finally, Section 6 brings a discussion of the results and our intentions for future work.

2. Background

According to [Liu 2009], the performance of focused web crawlers is measured with a set of metrics based on the similarity of the collected pages and the topic of interest. Such metrics include the *Average Precision*, which is the average of the similarity of all collected pages with the topic of interest, and the *Harvest Rate*, that can be formalized as the ratio between the number of relevant pages collected (i.e., the ones whose similarity exceeds a certain threshold) and the total number of collected pages.

The aim of focused crawling algorithms is to visit relevant pages first. To better accomplish this task, most of the state-of-the-art algorithms take into account the radius one hypothesis [Chakrabarti 2002], which states that an outlink of a page related to a subject is more likely to belong to the same subject than a random web page. We estimate the relevance of a page by calculating its cosine similarity with a query, representing them both in a vector space model.

To calculate the similarity between two documents in the vector space model , we set each weight of every vector corresponding to a document as the IDF (*inverse document frequency*) of a term multiplied by its normalized frequency. The IDF of a term expresses how rare t is in a collection C and is formalized as $IDF(t, C) = \log \frac{\|C\|}{df(t,C)}$, where $df(t, C)$ is the number of documents of C containing t.

One of the simplest algorithms for focused crawling is the Best-First Search (BFS), as described in [Liu 2009]. It uses the radius one hypothesis to guide the crawl, organizing the URL frontier prioritizing the outlinks of the most relevant pages. At each iteration, the page pointed by the URL located at the head of the queue is collected. A slight variation of this method is the *Best-N-First*, that collects the N first URLs in the queue, instead of just the first.

3. Related Work

In [Granada et al. 2012], a methodology for the acquisition of comparable corpora exploiting commercial web search engines is proposed. Their approach relies on the availability of a multilingual ontology which is used to identify the important concepts in the domain of interest. The ontology labels are combined in groups. Each of these groups is submitted to a search engine and the documents corresponding to the returned URLs are retrieved and parsed. An interesting particularity of this work is that the ontology labels can be multiword expressions, which tend to be less polysemic than single terms and, as a result, have a better expressive power. A potential limitation, however, is the dependency on a web search engine, which may lead to problems if the search engine changes the way their services are accessed, by limiting the number of allowed queries, charging fees for the services or simply changing the way requests must be done.

[Talvensaari et al. 2008] propose the use of focused web crawlers to acquire comparable corpora in order to train domain-specific MT systems. The crawling method is divided into two stages. First, queries with keywords related to the target domain are manually submitted to a search engine. Then, fifty more queries are built, using the most frequent words in the result pages. The results of these later queries are scored according to their frequency and rank inside a query and are grouped by host. The second stage is the crawling itself. The seed pages are the URLs with the highest scores in the hosts whose pages had the largest sum of scores. The score given to each page in the frontier are measured by the ratio of relevant words (words present in a query, built with the same terms used to build the previous fifty queries) in the anchor text where the link was found in the pointer page and in the set of pages belonging to the same host as the pointer page. The corpus obtained at the end of the process was used to train automatic MT systems. In their experiments, some texts related to the domain of interest were translated using the proposed approach and their results overcame the ones from the baseline translator, trained with a larger corpus but on a generic domain.

[Uszkoreit et al. 2010] present a methodology for using highly heterogeneous corpora for training statistical MT systems. The main goal is to identify document-pairs in which one is a translation of the other. For this, they are initially translated into a common language, using a baseline MT system. Then, two sets of n-grams are extracted from every document. The first is called the set of matching n-grams and is used to build a list of candidate document-pairs. Two documents are put in this list only if the number of common matching n-grams are equals to or higher than a certain threshold. The second is a set of lower order n-grams and is used to measure how similar are two documents from a candidate document-pair. The similarity between two documents is measured by using only the IDF feature of the document vectors. Finally, sentences are aligned from every document-pair and they are used to train a MT system. In the experiments, they used around one billion documents as input for their method.

The work of [Achananuparp et al. 2008] brings a comparison of many different metrics for sentence similarity. Those metrics can be divided in three main groups. The simplest group contains metrics that consider only the number of words shared by two sentences. The second one is composed by metrics based on the classic TF-IDF model. The third, and more complex one, is the group that contains language based measures, depending on resources that indicates semantic relations between terms. All those metrics are used to calculate similarity between sentences and the results indicate that although it does not always well evaluated with recall, the classic TF-IDF model usually provides a good precision.

Our work has similarities with [Talvensaari et al. 2008], since we also use corpora collected with focused crawling to train MT systems. Although we are aware that using search engines may be a good alternative to collect comparable corpora, we intend not to depend on their availability. We make use of the strategy of [Uszkoreit et al. 2010] by making an initial translation in order to have all our sentences written in the same language, for us to be able to compare them with the classic TF-IDF model. Details of our proposal are described in the Section 4.

4. Crawling for Comparable Corpora

This section details our focused crawling algorithm. Also, we describe how we preprocess and use the collected corpora to train a MT system. The experimental details are on the Section 5.

We developed an extensible focused crawler which, from a given set of seed pages and a set of relevant pages, used as reference to assess relevance, seeks to fetch other relevant pages from the domain and languages of interest. The crawler has a URL manager, which maintains a list containing all the already visited URLs and selects which of them should be visited next, according to a score assigned to each URL. Hence, it is only necessary to implement new ways of assigning scores to URLs to be able to compare the differences between two or more distinct crawling strategies.

Figure 1 summarizes the crawler structure. First, the *main* method sends a configuration file to the *Config* class to properly load all user defined options, which include definitions of parameters like n-gram size, the crawling algorithm, the set of seed pages, and locations of files to load features such as the IDF vectors. Then, it starts processing URLs received from the *URL Frontier Handler*, that sorts its URL frontier following the algorithm specified by the user, and sends the contents of the documents and the URLs to the crawler. These will be to managed according to the used method. The *Robots Analyzer* reads the $robots.txt$ file, present in the root of the host server, to prevent any disallowed URL from being added to the URL frontier. The score assigned to every document is computed by the *URL Frontier Handler*. Normally, it just calls the *Document Similarity* method, but it can be easily overwritten, in order to change the crawler's behavior. The *Document Similarity* method uses the standard cosine similarity measure between a document and a given query to estimate the relevance of every document. It is important to note that the relevance and the score are not necessarily equal, since the relevance is always calculated by the classic TF-IDF cosine similarity between the document and a query, while the score can be anything the user finds convenient. The *HTML Parser* is the module responsible for connecting with the servers, parsing the HTML documents, storing the texts in a repository, and sending them as a set of tokens to the *main* class.

All focused crawling algorithms that we have fully implemented are based on the *Best-N-First*. While a document is being parsed, every term found in it is stemmed and added to a vector that represents it. When the parsing finishes, the cosine similarity between this vector and a query is calculated and all the outlinks are added to the URL frontier, using this similarity as the score. We also implemented another algorithm that is slightly different from the *Best-N-First*, in what it uses n-grams to compose its vectors, instead of single terms.

After the crawling stage, the gathered corpora are pre-processed as follows. First, the corpora are split in sentences, which are, then, translated into a common language by a rule-based translation system. Then, we use the TF-IDF model to find similar sentence pairs that were originally written in different languages. The original, untranslated, text of similar sentences are used as input to train a statistical MT system.

Our main goal is to compare the quality of the focused crawling algorithms, measured with usual metrics, such as the harvest rate, with the quality of the lexical resources that can be extracted from the collected corpora, to establish a correlation between both. Besides, we propose to use the quality of these resources, measured by the accuracy of the generated translations, as an extrinsic measurement to evaluate a focused crawler.

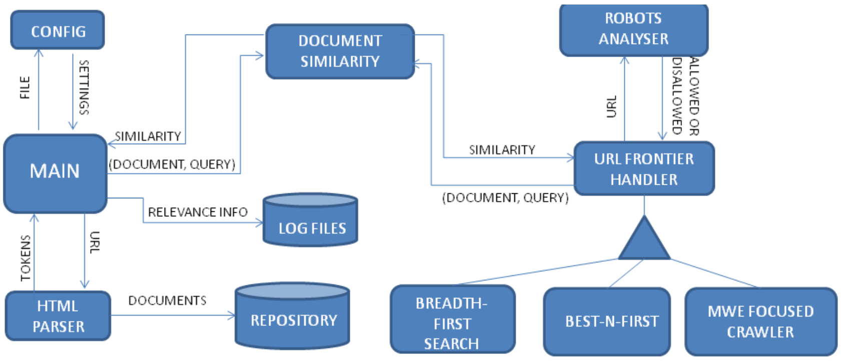

Figura 1. Basic crawling procedure

5. Preliminary Experiments

The goal of the experiment was to use the collected corpora to translate texts between Portuguese, English, and French on the *genetics* domain. For the crawling stage, the linguistic resources used were generic corpora to compute IDFs for every term n-gram present in them. Newspaper collections from the years 1994 and 1995 for each of the languages were used. For English, the Los Angeles Times and Glasgow Herald were used; for French, we used texts from Le Monde and from the Swiss Document Agency. All these collections are available in the CLEF Test Suite, which can be purchased from ELRA[1]. For Portuguese, we used texts from Folha de São Paulo[2]. Due to memory limitations, we pruned all terms and n-grams that appeared in fewer than five documents.

We compared the *Best-50-First* (*Best-N-First* with $N = 50$) using single terms and using bigrams and trigrams. For every crawling algorithm, our stopping criteria was to reach 10000 successfully collected pages. The next stage was to split the output corpora into sentences. To perform this task, we used the Python *Natural Language Toolkit* [Loper and Bird 2002] sentence splitter for all three languages. Sentences containing less than 3 or more than 22 words were pruned. We did this to ignore stand alone terms or expressions, such as titles or menus, and cases where the sentence splitter grouped sentences and expressions together. We chose 22 to be the maximum number of words allowed in a sentence because of the analysis done in [Granada et al. 2012], who computed the average number of words in a comparable corpus. In Portuguese, they found 20.07 words per sentence, while for French and English, the numbers were 15.91 and 14.15, respectively.

To perform the sentence alignment, we first use Moses, which is an open-source statistical MT toolkit, that offers several tools for preprocessing, training a translator with parallel, data and translating sentences [Koehn et al. 2007]. We used a Moses translator trained with parallel sentences from the Europarl [Koehn 2005], to translate all sentences written in Portuguese and French to English. In order to avoid comparing all possible sentence-pairs, we index our sentences with Zettair[3], a search engine that allows the user to build indexes and make customized queries. To find what sentences are similar to a given sentence, we send her as a query to Zettair, that searches for similar sentences in an index that was built with all sentences that were originally written in one of the other two languages. We also set the search engine's ranking function to be the cosine similarity metric. Finally, we use the original forms of all pairs of similar sentences to train Moses.

To reach our main goal, which is to establish a correlation between the focused crawling performance and the quality of the lexical resources extracted from the collected corpora, we must know results of both measurements. We estimate the focused crawling performance using the standard metrics harvest rate and average precision. To evaluate the quality of the lexical resources, we generate translations about the genetics domain with the MT system trained with our comparable corpora. The translations are judged by humans that speak both source and target languages and the quality of the resources is measured from these judgments.

We have fully ran our experiments only until the crawling stage. The results indicate that the *Best-50-First* with unigrams outperforms the one with trigrams, which showed better results than bigrams. Although the sentence alignment and translation stages are still not fully implemented and evaluated, we expect the translator trained with corpora collected with the first algorithm to outperform the other ones.

[1] http://www.elra.info/

[2] Available from http://www.linguateca.pt

[3] http://www.seg.rmit.edu.au/zettair/

6. Conclusion

The aim of this work is to evaluate the quality of focused crawling algorithms and the quality of the lexical resources extracted from the respective collected corpora, in order to find a correlation between both and propose to use the quality of the resources as an extrinsic metric to evaluate focused crawlers. We have already implemented and experimented *Best-50-First* algorithm using single terms, bigrams and trigrams. Besides, we are still implementing a focused crawler that uses Hidden Markov Models to learn user browsing patterns [Liu et al. 2006] and developing another one that takes advantage of the expressive power of multiword expressions. Finally, to be able to establish a correlation between the measurements, we still need to manually judge a considerable quantity of generated translations between all language pairs.

Acknowledgements: The first author receives a scholarship from CNPq. This work was partially funded by CNPq (projects 478979/2012-6 and 480283/2010-9) and CAPES-COFECUB (project 707/11).

Referências

Achananuparp, P., Hu, X., and Shen, X. (2008). The evaluation of sentence similarity measures. In *Data Warehousing and Knowledge Discovery*, pages 305–316. Springer.

Chakrabarti, S. (2002). *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann.

Granada, R., Lopes, L., Ramisch, C., Trojahn, C., Vieira, R., and Villavicencio, A. (2012). A comparable corpus based on aligned multilingual ontologies. In *Proceedings of the First Workshop on Multilingual Modeling*, pages 25–31. ACL.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. ACL.

Liu, B. (2009). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer.

Liu, H., Janssen, J., and Milios, E. (2006). Using HMM to learn user browsing patterns for focused web crawling. *Data & Knowledge Engineering*, 59(2):270–291.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. ACL.

Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., and Laurikkala, J. (2008). Focused web crawling in the acquisition of comparable corpora. *Information Retrieval*, 11(5):427–445.

Uszkoreit, J., Ponte, J. M., Popat, A. C., and Dubiner, M. (2010). Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1101–1109. ACL.