Named Entity Translation using Anchor Texts

Wang Ling[1][2], *Pável Calado*[1], *Bruno Martins*[1], *Isabel Trancoso*[1], *Alan Black*[2], *Luísa Coheur*[1]

L^{2}F Spoken Systems Lab, INESC-ID, Lisboa, Portugal[1]
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA[2]
{wang.ling,pavel.calado,bruno.martins,imt,luisa.coheur}@inesc-id.pt
awb@cs.cmu.edu

Abstract

This work describes a process to extract Named Entity (NE) translations from the text available in web links (anchor texts). It translates a NE by retrieving a list of web documents in the target language, extracting the anchor texts from the links to those documents and finding the best translation from the anchor texts, using a combination of features, some of which, are specific to anchor texts. Experiments performed on a manually built corpora, suggest that over 70% of the NEs, ranging from unpopular to popular entities, can be translated correctly using sorely anchor texts. Tests on a Machine Translation task indicate that the system can be used to improve the quality of the translations of state-of-the-art statistical machine translation systems.

1. Introduction

Named Entity (NE) translation plays a major role in many Natural Language Processing Applications. One example of such an application is Statistical Machine Translation, where many NE translations cannot be found in the parallel corpora used to train the translation models, so translation dictionaries of NEs are employed. Other applications include cross-lingual Information Retrieval and cross-information Question Answering, where higher quality NE translations contribute to obtaining better results. Pre-compiled NE translation dictionaries, such as the Chinese-English Translation Lexicon from LDC (Linguistic Data Consortium), are one possible source for accurately obtaining these translations, but these are hard to find for many language pairs and their coverage is limited unless they are updated frequently, as new NEs, such as movies and celebraties, emerge everyday. In this work, we present a web based approach for finding NE translations automatically by leveraging the text in web links, which we refer as anchor texts.

This paper is organized as follows: in Section 2, we describe some background about NE translation and in Section 3, we present some work on anchor texts. The description of the algorithm can be found in Sections 4, 5 and 6. Section 7 presents the results of the NE translation algorithm and follows with the analysis of the same results. In section 8 we study the impact of our system in machine translation systems. Finally, we conclude in Section 9 and propose some future work.

2. Named Entity Translation

An algorithm for extracting named entities translations from parallel corpora, with high accuracy is illustrated in [1]. The main drawback of this method is that it relies on parallel corpora, which is a scarce resource, making this method inadequate for finding NE translations of not frequently occurring NEs. Although, there are algorithms to extract parallel corpora from the World Wide Web [2], the amount of parallel documents that can be extracted constitute only a very small portion of the Web.

Many successful efforts have been invested in the automatic NE translation using Web resources. These methods include a methodology for searching web documents where the untranslated or source NE occurs. The algorithm described in [3] uses words relevant to the NE, translates them to the target language, allowing these to be used as hint words to improve accuracy and reduce ambiguity. For instance, the word "Python" can be a type of snake, the programming language, the movie, or the name of a revolver gun (Colt Python). If the document where the word Python was inserted contained many instances of "programming" or "code", these can be translated into the target language and used in the query as disambiguation words.

After extracting the documents, the next task is to determine the correct translation from these documents. This is generally done by extracting various candidates, filtering the candidates and scoring these based on confidence scores of each candidate. The work done in [4] and [5] uses phonetic transliteration, for Chinese, to determine the translation candidate that has the highest phonetic similarity with the NE in the source language. These methods work well with named entities that are translated phonetically, but this is not always the case. Some entities are translated semantically, for instance, "The Day After Tomorrow" is translated to "O Dia Depois de Amanhã" in Portuguese and "后天" in Chinese. For these types of NEs, it is more adequate to explore semantic similarities between the NEs in different languages. The work in [6] combines semantic and phonetic similarities to achieve an accuracy of 67% for rarely occurring NEs.

A different approach [7] explores patterns that occur in documents that may indicate a possible translation of a named entity, such as “后天(The day after tomorrow)”, from which a person should be able to deduce that the NE within the brackets are the translation of the NE that immediately come before. This kind of deduction is learned using a manually annotated corpora, where the NEs and their translations are identified.

3. Anchor Text

Anchor texts are defined as the visible, clickable text in a hyperlink. They have been used successfully on several applications including query refinement [8], where their usage produced better results compared with methods that extracts refinement terms for a query using a document collection. The authors of this work, identify several benefits of anchor texts. One benefit is the fact a collection of anchor texts contain much less text than a collection of documents, therefore, processing such a collection is faster.

In our work, we exploit a useful property of anchor texts that is the fact that they are a very succinct description of the target web page, which in turn facilitates retrieval of the correct translation. For instance, many links that are linked to a web page about a famous personality, will have that person’s name and possibly some other words. On the other hand a web page about that person is likely to contain a few paragraphs of texts giving a detailed description about that person. When extracting NE translations, having a large amounts of text besides the NE translation introduces many more incorrect candidates, which makes the task of retrieving the correct translation harder. The wikipedia entry for “Bill Gates” contains approximately 5000 words of content, yet, the name only appears approximately 40 times in the document, which is near 1% of the document. Furthermore, many other named entities are mentioned in the document such as Microsoft, Paul Allen, Harvard and General Electric. On the other hand, the set of links to that web page, which we will call anchor set, contains approximately 1500 links, and the text of 900 links contains the word “Bill Gates”. This does not take into account misspellings, translations of the name and other designations such as “Mr Gates” and “William Henry Gates III”. Furthermore, around 800 of the 900 links had no additional words beside the name, while the rest contained at most 4 additional words, such as “wikipedia”.

The work presented in [9] leverages anchor texts in order to extract term translations for query translation. This work takes advantage of the fact that links to the same web page generally share similar information. Thus, if the text in those links are in different languages, they might contain entities that are translations of each other. To translate an NE in a language (source) to another language (target), their system processes an anchor set, which is built by extracting all links to a web page with at least one link with the source NE, and extracts that term’s translation based in the following joint probabilistic model:

$$P(s \leftrightarrow t) = \frac{P(s \cap t)}{P(s \cup t)} = \frac{\sum_{i=1}^{n} P(s \cap t \cap u_i)}{\sum_{i=1}^{n} P((s \cup t) \cap u_i)} \quad (1)$$

The equation above estimates the similarity between the source term s and target translation t. This model makes the assumption that anchor texts in different languages that co-occur frequently and occur rarely separately are likely to be translations of each other in the respective language. It also considers web pages $U = u_1, u_2, ...u_n$ with higher authority more reliable. The actual implementation further assumes that s and t are independent given u_i, producing the following equation:

$$P(s \leftrightarrow t) = \frac{\sum_{i=1}^{n} P(s|u_i)P(t|u_i)P(u_i)}{\sum_{i=1}^{n} P(s|u_i)+P(t|u_i)-P(s|u_i)P(t|u_i)]P(u_i)} \quad (2)$$

The values of $P(s|u_i)$ and $P(t|u_i)$ are estimated by calculating the faction of the u_i’s in-links (links to u_i) containing s and t, respectively, over the set all in-links to u_i. The authors discuss that their method is able to find translations of NEs that appear frequently in the anchor sets, but is not as effective with rarer terms. Furthermore, the web page mining was done using a crawler over web pages in Taiwan, which only extracted 1,980,816 million Chinese web pages, and they refer that a increase in the anchor set might improve the performance of the system, but might also result in more noisy data.

In our work, we design an NE translation system using anchor texts, but we suggest a different approach that is able to overcome some of the shortcomings of the work presented above.

First of all, we consider that equation 2 is too restrictive for translation extraction, specially for rarely occurring terms or NEs. The main reason is that if the source term is not contained in the in-links to a web page, all other in-links to that page are discarded. For frequently occurring terms, we can generally extract a large anchor set and can afford to discard some in-links, but for rarely occurring terms, discarding these will considerably reduce the probability of finding the translation in the resulting anchor set. For instance, the Chinese translation for the NE “George Kaiser” and “Alice Walton” is only present in one in-link in the anchor set generated by our system. Using the criteria described above, these links would have been filtered out because no in-link from the web page had the NE in English. Furthermore, web pages about rare terms contain a small sample of in-links. Most web pages for the 2 NEs above had less than 2 in-links, so it is unlikely that one in-link contains the term in English and another link with it in Chinese. Thus, we believe that the probability estimation for $P(s|u_i)$, would be more accurate by also considering the content in u_i.

Secondly, we find that due to the small scale of the anchor set used in the work above the results in the system described above cannot be properly compared to NE translation systems that leverage search engines that index billions of web

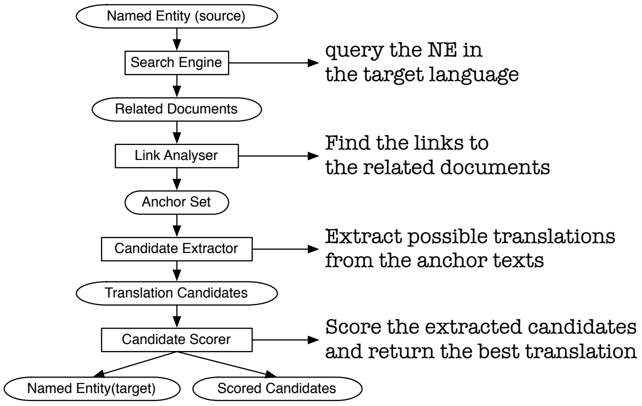
Named Entity (source)
Search Engine
query the NE in the target language
Related Documents
Link Analyser
Find the links to the related documents
Anchor Set
Candidate Extractor
Extract possible translations from the anchor texts
Translation Candidates
Candidate Scorer
Score the extracted candidates and return the best translation
Named Entity(target)
Scored Candidates

Figure 1: Overall structure of the NE translation algorithm.

pages. Taking this into account, we shape our the architecture of our system to leverage online resources, allowing a larger application scale.

Finally, using co-occurrences frequency to find the correct translation of an NE among the candidates works well for terms that occur often. However, for terms that occur rarely, frequency is not enough to find the correct translation. In the case the correct translation only appears once, it will have an equal or lesser probability in contrast with all other candidates. We define our system to use multiple features to improve the accuracy of our system.

4. NE Translation Extraction Algorithm

Figure 1 illustrates the overall structure of our system.

Given an NE e_s in the source language s that we want to translate to e_t in the target language t, we first retrieve a set of web pages $d_1, d_2, \ldots, d_k \in D$, whose in-links might contain the translation for that named entity. This is done using web queries, similarity to the work in [3] and [4]. We use the Google Custom Search API (code.google.com/apis/customsearch/) and perform a query using e_s as the query word. Furthermore, we also set the API to give higher priority to web pages in the target language.

Then, we retrieve the set of anchor texts $a_1, a_2, \ldots, a_m \in A$ for each element d_i^e in D_e. We denote A_{d_i} as the set of anchor links that are linked to d_i. This set is extracted using the SEOmoz's Open Site Explorer (www.opensiteexplorer.org), which automatically extracts the anchor in-links to a specified d_i. Furthermore, it also provides some useful information such as the authority of target web page, based on link analysis.

After retrieving A, we need to extract the possible translations $e'_t \in T$ for the source NE e_s, and find the best translation $\hat{e}_t$ such that:

$$\hat{e}_t = argmax_{e'_t \in T} P(e'_t|e_s) \tag{3}$$

Where $P(e'_t|e_s)$ is the probability of e'_t being the translation for the source NE e_s. The set of candidates T is and is defined by all possible n-grams from text of all anchor links in A. Generally, performing the search over all possible values of T is intractable due to the large number of possible n-grams in web pages, so the set of T is limited to the detected NEs in those pages. However, the number of candidates in the text in anchor links is tractable, since they are an order of magnitude smaller than web pages.

We estimate $P(e'_t|e_s)$ using a linear combination of various features $f_1, f_2, \ldots, f_n$ and weighted by $w_1, w_2, \ldots, w_n$. Thus, we the best translation is given by:

$$\hat{e}_t = argmax_{e'_t \in T} \sum_{i}^{n} f_i(e'_t|e_s) w_i \tag{4}$$

5. Features

As previously stated, the probability of a translation candidate e_t is given by a combination of weighted features. We define the feature relative frequency $freq$ as:

$$freq(e_t|e_s) = \frac{C(e_t)}{\sum_{e_t' \in A} C(e_t')} \tag{5}$$

, where $C(e_t)$ is defined by the number of times e_t appears in A.

Anchor texts such as "click here" and "in english" appear frequently in every anchor set and are evidently not good candidates for the translation. Thus, we add the feature global frequency g_freq to act as a counterweight to the frequency. The g_freq calculates the relative frequency of each candidate as $freq$, but uses a general anchor set A_g, that is not specific to e_s. The A_g is built by merging the anchor sets for all entities in the training/test corpora. This means that anchor texts such as "click here" that generally occur frequently in all anchor sets will have a high value for g_freq, and chances that these will be chosen become lower.

In many languages, such as Portuguese and English, NEs contain approximately the same number of words. For instance the movie "Harry Potter and the Sorcerer's Stone" is translated into Portuguese as "Harry Potter e a Pedra Filosofal", and both contain 6 words. Thus, the word distance feature $dist$ is defined as:

$$dist(e_t|e_s) = |length(e_t) - length(e_s)| \tag{6}$$

While named entities in different languages do not always have similar sizes, this feature can eliminates many incorrect candidates that are unrealistically large.

We also use Named Entity Recognition as a feature, since we do not filter out candidates that are not recognized as NEs, because the system would be dependent on the correct detection NEs. We use 2 features to detect named entities. First, we use the Stanford Named entity recognizer to detect NEs. We also look for patterns such as words that occur in quotes, such as "the movie "Lord of the Rings"". Another example for Chinese is the fact that peoples names that are translated from English are sometimes separated by “ · ”, for instance, “Harry Potter” is written as “哈利 · 波特”, where “Harry” is translated to “哈利” and “Potter” is translated to “波特”. These features are defined as boolean feature, telling whether a candidate is considered a named entity.

We use the co-occurrences feature co_freq, which counts the number of co-occurrences of e_s and e_t occur in the anchor texts linking to the same web page. This is defined as:

$$co\_freq(e_t|e_s) = \frac{C(e_t, e_s)}{\sum_{e_t' \in A} C(e_t', e_s)} \tag{7}$$

While this is similar to the $freq$ feature, since the anchor set A is extracted from web pages using e_s as a query, the documents returned by the search engine might contain web pages that are not entirely about the e_s but about a related topic. For instance, a query for a movie by name might return web pages about its directors, therefore anchor links to those pages are more likely to have the name of the director, and might generate incorrect candidates. Thus, if the e_s occurs frequently in the anchor texts linking to a web page, it is more likely that the web page is more focused on e_s and nota related topic.

We also found that, in many web documents, authors tend make links that only contain the NE. To take advantage of this property, we define the link occurrences feature $link\_freq$, which is the relative frequency e_t is the whole link in A.

6. Translation Candidate Scoring

Given the features, we use Linear Regression to train the weights for each feature, using a training set of source NEs and their translations.

Based on the reference translation r and our system's hypothesis h, we test 2 different score functions, which we want to maximize.

The first scorer is simply given by:

$$S(h, r) = \delta(h, r) \tag{8}$$

where we $\delta(h, r)$ is the Kronecker delta and returns 1 when $h = r$ and 0 otherwise.

We test another scorer that calculates the similarity between the words in the r and h, defined by:

$$S(h, r) = \frac{H \cap R}{H \cup R} = \frac{\#(H \cap R)}{\#H + \#R - \#(H \cap R)} \tag{9}$$

where H and R are sets with all the words in h and r, respectively, and $\#H$ and $\#R$ are the number of words in the respective sets. We use this second scorer is due to the fact that some candidate translations might be partially correct, and might be used to estimate the optimum weights for the features more accurately.

In the following sections we will refer to the system trained using the first score function as classification, since the scorer divides the samples into two classes, 0 and 1, and the system trained with the second score function as regression.

Experiment	Training	Test
Countries English-Chinese	56 countries + 50 people	168 countries
Countries English-Portuguese	56 countries	168 countries
People English-Chinese	56 countries + 50 people	48 people

Table 1: Datasets description.

7. Named Entity Translation Results

To evaluate our method, we manually built various test sets to test our system. The following subsections will describe the process used to gather the test corpora, the results that were obtained and our analysis of these results.

7.1. Corpora

We processed the wikipedia entry with the list of countries ordered by population (en.wikipedia.org/wiki/List_of_countries_by_population). The entry is translated in various languages and a simple script was created to retrieve the list in different languages in the same order. As a result, we obtained a dictionary of 224 countries. Using a similar approach we also built a dataset with peoples names for the Chinese-English pair by parsing the Chinese version of the Forbes list of billionaires, which contains 100 entries with both the English and Chinese names of the 100 persons.

7.2. Results

We conducted 3 experiments, which are illustrated in Table 1. The first 2 experiments were conducted for the countries dataset, for the English-Chinese and English-Portuguese language pairs, while the third experiment was conducted using the peoples data set for the English-Chinese pair.

For each experiment, we calculate the percentage of the test set that was translated correctly. We check the percentage that was in the anchor set to discern whether incorrect translation is derived from the NE being not present in anchor set extraction or from the incorrect identification of the candidate. These results can be found in table 2. Finally, we generate the list of candidates, ordered by score, and plot the number of correct translations found in the top-n candidates up to 20. These can be found in figures 2, 3 and 4.

7.3. Accuracy and Coverage

We evaluate our system in two levels. We evaluate how well our system retrieves documents that contain the translation of each NE by defining coverage as the percentage of the NEs that can be found in the anchor texts. We also evaluate whether we can identify the correct answer from the re-

Experiment	correct	exists
Classification		
Countries English-Chinese	**71%**	94%
Countries English-Portuguese	**62%**	89%
People English-Chinese	**22%**	31%
Regression		
Countries English-Chinese	70%	94%
Countries English-Portuguese	**62%**	89%
People English-Chinese	20%	31%

Table 2: Results for each experiment. The "correct" column shows the percentage of NEs in the test set that were translated correctly and the "exists" column displays the percentage of the NEs that exist in the anchor set for that NE.

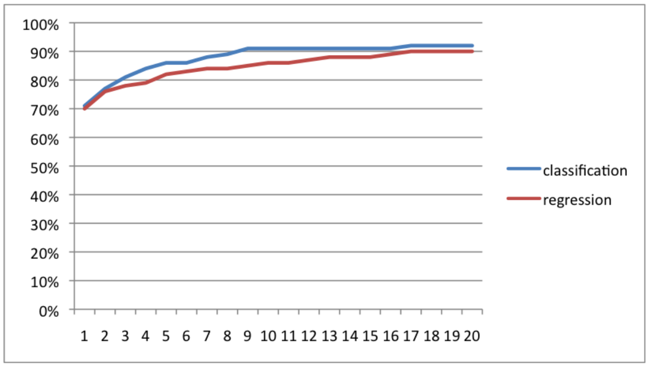

Figure 2: Plot with the percentage of correct terms (Y-axis) that were included in the top n candidates (X-axis) for the countries corpus(EN-CN), ordered by score.

trieved anchor texts by defining accuracy as the percentage of correctly translated NEs from those that exist in the anchor set.

The results for the countries test set for the English-Chinese pair are considerably higher than the other 2 experiments. In fact, if we take into account that we had a coverage of 94%, the accuracy of the translation, using classification, would be 76%.

In contrast, the accuracy of the English-Portuguese experiment was evidently lower, achieving only an accuracy of 70%. This can be associated with the fact that the Portuguese setup did not use a NE recognizer , which is a extremely good indicator to exclude many incorrect candidates. However, this experiment attains a coverage of 89%, which is nearly as high as the English-Chinese experiment, even considering that the magnitude of web documents in Chinese far is superior than those in Portuguese. This is partially derived from the fact that 55 of the 168 samples in the countries in the testset in English are written identically in Portuguese, therefore, finding web pages about those NEs in Portuguese simple. Another reason for this, is that many words such as India, Indonesia are translated to Portuguese by simply adding a accented vowel to form Índia and Indonésia. These are considered orthographic errors and corrected automatically, hence, web pages about these NEs will also be easily found by the search engine. In the test set, there are 18 samples with fall in this category. Finally, there are words that are very similar in Portuguese and English such as Brasil and Brazil, which are also corrected. We cannot find the number of samples in this category since the exact specifications of the googles search engine's are not known. Taking all these together, we estimate that web documents for nearly half of the NEs in the test set can be trivially retrieved by the search engine. If we only use 73 NEs that fall in the first 2 categories, we get a coverage of 98%, with only one NE that is not found, which is "Côte d'Ivoire". In contrast, using the remaining 95 NEs as a test set would result in a coverage of 82%.

The experiment for the English-Chinese language pair for people has the very low coverage of 31%, while the maintaining a considerably high accuracy of 70%. Moreover, we can see in figure 4 the all the correct translations are found within the top 5 candidates, which does not happen in the other 2 experiments. This high accuracy is related to the fact

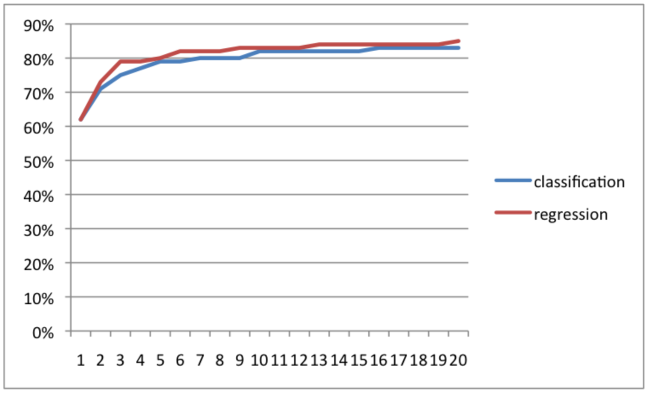

Figure 3: Plot with the percentage of correct terms (Y-axis) that were included in the top n candidates (X-axis) for the countries corpus(EN-PT), ordered by score.

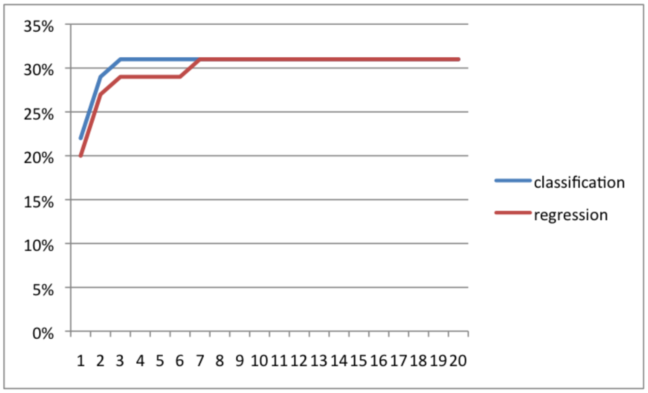

Figure 4: Plot with the percentage of correct terms (Y-axis) that were included in the top n candidates (X-axis) for the people corpus(EN-CN), ordered by score.

that considerably large numbers of people's full names are separated in chinese with the token “ · ”, such is the case with “埃克 · 巴蒂斯塔” (Eike Batista). This makes identifying NEs much easier.

Regarding the low coverage, by looking at the best translation candidates given by our system and comparing these to the reference, we noticed that part of the names given by our system but were not equal to the reference were actually alternative designations of the person's names. For instance, the NE “William Gates III”, the reference translation is “比尔盖茨”, while our translation is “比尔盖兹”, differing only in 1 character. The characters that differ, “茨” and “兹”, have the same phonetic transliteration. While the reference might be more correct, the translation of our system is also understandable, and probably more widely used on the Web.

7.4. Classification vs Regression

We can see from figures 2, 3 and 4, that using similarity as a score in the training step produces better results for the English-Portuguese language pair and worse results from the other pairs. By looking at the test and training sets we think we can induce some causes for these results. Many countries written in portuguese have a reduced designation. For instance, “Estados Unidos da America” and “República Popular da China” can be reduced to “America” and “China”, which are found more often in the anchor set. If we use the scoring method in the classification, the reduced designations will be considered negative samples, even though these are actually positive samples. In the regression, since we are using a similarity based scorer, those samples are considered partly correct, which, in turn, produce more accurate results.

On the other hand, when the target language is Chinese, the opposite effect occurs. This effect roots from the fact that many countries and peoples names in Chinese have common characters. The most evident one is the character “国”, which occur very frequently in countries such as “美国” (America)，“中国” (China)，“法国” (France)，“德国” (Germany)，“英国” (England) . Furthermore, NEs that are phonetically translated generally use the same set of characters for the translation of each phoneme. For instance, the character “卡” is used in the translation of “Carlos”, “Karl” and “Deripaska” to translate the “kA” phoneme. According to the expression used to calculate the similarity score, many incorrect NEs that are found in the anchor set will have considerable score, which in turn generates noise in our model during the training step. For countries such as “法国” (France), if we find “美国” (America) in the anchor set, the respective candidate will have a 0.33 score, which is a considerably elevated score for a negative example.

We attempt to demonstrate this by processing the countries training and test data, and calculating the number of times characters repeat, in different NEs, after their first occurrence, which means that a character that occurs 3 times will be repeated twice and a character that occurs 9 times will have 8 repetitions. For the Chinese data, 545 of 811 (67%) characters are repeated, while in the Portuguese data, only 41 of 306 (13%) words re-occur. The 5 words or characters that occur the most for each language can be found in table 3, where can see that the number of the re-occurrences in Chinese is a order of magnitude higher than in Portuguese.

Word / Character	occurrences
Chinese	
亚	41
斯	27
尼	20
拉	18
巴	18
Portuguese	
Ilhas	10
do	7
e	7
República	6
São	4

Table 3: Number of occurrences of the words that have the highest occurrences for Chinese and Portuguese.

7.5. Anchor Text vs Web Document

We now compare the characteristics of the web documents and the text in the links pointing to them in terms of size, coverage, processing time and the ratio between NE translations with the size of the documents. We use the test corpus of the people's names corpora for this analysis, because it has a relatively low coverage so we want to analyse whether web documents can produce better results.

The web documents were striped from their HTML tags, and the statistics were extracted and are presented in Table 4. The size is expressed as the number of Chinese Characters in the whole collection of documents or anchor texts (Total Size), and also the average size in these documents (Avg Size), the Coverage is expressed as the percentage of NEs in the test set that were found. The Ratio is the percentage between the correct translations in the documents and the total number of characters in those documents. Finally, the Processing Time is the time that took to produce these results.

We can clearly see that the total size and average size of the Web Documents and the Anchor Links are in different orders of magnitude, in fact, many web documents that were retrieved were larger than the whole anchor set. This, in turn, leads to the large time needed to process web documents, whereas processing the anchor set is done almost instantly.

In terms of coverage, the retrieved set of web documents contain the translations of nearly all the samples in test set, yet the ratio between the characters that are the correct translations and the remainder of the text is much lower than the same ratio for the anchor texts. This indicates that using web documents would achieve a high coverage, but it will be much harder to find the correct translation within the can-

Measure	Anchor Texts	Web Documents
Total Size	13900	41191507
Avg Size	24	40 583
Coverage	15(31%)	46(96%)
Ratio	0.1%	0.00336%(1386)
Processing Time	6 secs	17 mins + 31 secs

Table 4: Comparison between Anchor Texts and Web Documents in terms of Total Size, Average Size, Coverage, Ratio and Processing Time.

Entity	Type	Proposed	Reference
Luxemburg	LOCATION	卢森堡	卢森堡
Moore	PERSON	穆尔	穆尔
Charleston	LOCATION	查尔斯顿	查尔斯顿
Spock	PERSON	美国	斯波克
Seiji Ozawa	PERSON	小泽征尔	小泽征尔

Table 5: Untranslated NEs in the Dialog test corpora and the proposed translations from our proposed NE translation system.

didates. On the other hand, using anchor texts can achieve a higher accuracy, due to the relatively small amount of incorrect translations candidates, but it would be much harder to retrieve anchor texts with the correct translation.

Based on this results, we believe that the next logical step is to invest in the anchor set retrieval steps, rather than in the candidate scoring steps. For instance, we search the untranslated NE and restrict the results to the target language. This means that only web pages with the untranslated NE will be used, which constitute only a small portion of the web pages that are might contain useful anchor links. Examples of other methods to relax this restriction is to use keywords that are related to the NE in the source language and querying using those that are translatable. For instance, if we do not know the name of an actor we could search for the movies where he stars in. Also, many web pages that are returned by Google Search Engine are not indexed by the SeoMOZ's Open Site Explorer. It is likely that the usage of additional resources to retrieve anchor texts from documents would lead to a improvement in coverage.

Combining anchor texts with the web documents is another promising expansion for our system. This could improve the results of state-of-the-art NE translation systems, since anchor texts have a higher ratio of the correct translations, these could be used as additional features to improve the accuracy of those systems. The only drawback is the fact the combined system will not have such low processing times.

8. Machine Translation Results

To evaluate the impact of our NE translation system on machine translation tasks, we apply our system to the IWSLT 2010 (http://iwslt2010.fbk.eu/) evaluation. Our experiments were performed over the dataset for the DIALOG task in the English to Chinese direction. The DIALOG corpus is a collection of human-mediated cross-lingual dialogs in travel situations. The training corpus for these tasks contains about 30K sentences. The development corpus contains approximately 200 sentences for the DIALOG corpus and the testset contained around 500 sentences. All the results were evaluated with 16 references using BLEU-4. The translation system used in our experiment uses the Geppetto toolkit [10] to train the translation models and the Moses decoder [11] to perform the translation. The system description can be found in [12].

A list of lexical gaps in the testset are extracted and the NEs are identified using the a the Stanford NE recognizer. 45 NEs where found and 5 were not translatable by the system. There were not many lexical gaps in the system, since the training corpora and test corpora were in the same domain. The untranslatable NEs that were found and the translations proposed by our NE translation system are illustrated in table5. Our system was able to find the correct translation for 4 out of 5 NE(80%), which is consistent with the values obtained with the testset of our NE translation system. It is also worth mentioning that the only NE that was not found was not in the anchor set for that NE. As for the translation quality, the insertion of the NE translations improved the results from 44.52 to 44.64. These results suggest that the NE translations using this system improves the results for machine translation.

9. Conclusions and Future Work

In this paper we presented a method to use anchor texts from web pages to obtain NE translations. A study was done, which shows that, in general, anchor texts contain a higher ratio between the number of times the NE occurs and the total number of words, in comparison with web pages. This reduces the entropy in our model, since the amount of incorrect NE translations decreases. Furthermore, anchor texts are also several orders of magnitude smaller than web documents and can be processed faster, or allow the application computationally expensive algorithms.

Two candidate scoring methods were tested. The first scorer returns 1 if the candidate is equal to the reference and 0 otherwise, while the second scorer uses a similarity measure between the candidate and the reference as the score. Tests on the data sets indicate that the second scorer can improve or deteriorate the models depending on the training set. If the reference entries contain large numbers of commonly used words/characters the similarity measure might consider entirely incorrect candidates as partially correct, introducing noise to the model. This is the case when the target language is Chinese. In the case of Portuguese this measure yields better models, since the number of common words is low.

The tests on the countries data set, yielded on the best conditions an accuracy of 76% and a coverage of 94% for the English-Chinese data set, and an accuracy of 70% and a coverage of 89%. The analyses of the people's names data set, showed that there is a need to consider alternative name translations when the target language is Chinese, since the sequence of Characters might differ depending of the author. Due to this problem, the coverage of this set is only 31%, but with an accuracy of 70%.

Tests on the IWSLT 2010 DIALOG task, showed an improvement of 0.12 points in BLEU. This roots from the correct translation of 4 out of 5 NEs that were, previously not present in the training corpora of the statistical machine translation system.

In the future, we plan to combine the anchor texts set in the links with the web documents they are linked to, which we believe that will improve the the coverage, since the anchor texts that are not found in the anchor set, can be found in the web documents and viceversa. The additional data from the web pages can also improve the accuracy of the system, by providing additional features for each candidate.

10. Acknowledgements

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also through projects CMU-PT/HuMach/0039/2008 and CMU-PT/0005/2007. The PhD thesis of Wang Ling is supported by FCT grant SFRH/BD/51157/2010. The authors also wish to thank the anonymous reviewers for many helpful comments.

11. References

[1] R. C. Moore, "Learning translations of named-entity phrases from parallel corpora," in *IN PROC. OF EACL*, 2003, pp. 259–266.

[2] P. Resnik and N. A. Smith, "The web as a parallel corpus," *Comput. Linguist.*, vol. 29, no. 3, pp. 349–380, 2003.

[3] Y. Zhang, F. Huang, and S. Vogel, "Mining translations of oov terms from the web through cross-lingual query expansion," in *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2005, pp. 669–670.

[4] L. Jiang, M. Zhou, L. feng Chien, and C. Niu, "Named entity translation with web mining and transliteration," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007, pp. 1629–1634.

[5] R. Sproat, T. Tao, and C. Zhai, "Named entity transliteration with comparable corpora," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ser. ACL-44. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 73–80.

[6] F. Huang, S. Vogel, and A. Waibel, "Improving named entity translation combining phonetic and semantic similarities," in *HLT-NAACL 2004: Main Proceedings*, D. M. Susan Dumais and S. Roukos, Eds. Boston, Massachusetts, USA: Association for Computational Linguistics, May 2 - May 7 2004, pp. 281–288.

[7] C. hao Yeh, W. chi Tsai, Y. chun Wang, and R. T. han Tsai, "Generating patterns for extracting chinese-korean named entity translations from the web," 2010.

[8] R. Kraft and J. Zien, "Mining anchor text for query refinement," in *Proceedings of the 13th international conference on World Wide Web*, ser. WWW '04. New York, NY, USA: ACM, 2004, pp. 666–674.

[9] W.-H. Lu, L.-F. Chien, and H.-J. Lee, "Translation of web queries using anchor text mining," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 1, no. 2, pp. 159–172, 2002.

[10] W. Ling, T. Luís, J. Graça, L. Coheur, and I. Trancoso, "Towards a general and extensible phrase-extraction algorithm," in *IWSLT '10: International Workshop on Spoken Language Translation*, Paris, France, 2010, pp. 313–320.

[11] H. Hoang, A. Birch, C. Callison-burch, R. Zens, R. Aachen, A. Constantin, M. Federico, N. Bertoldi, C. Dyer, B. Cowan, W. Shen, C. Moran, and O. Bojar, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 177–180.

[12] W. Ling, T. Luís, J. Graça, L. Coheur, and I. Trancoso, "Towards a general and extensible phrase-extraction algorithm," in *The INESC-ID Machine Translation System for the IWSLT 2010*, Paris, France, 2010, pp. 81–85.