

An Evaluation of Greek-English Cross Language Retrieval within the CLEF Ad-Hoc Bilingual Task

Polyxeni Katsioui, Theodore Kalamboukis
Department of Informatics
Athens University of Economics and Business
Athens 104 34

Abstract: This article describes an experimental investigation on the use of resources from the web on a common Natural Language Problem (NLP) problem that of Word Sense Disambiguation (WSD). In particular we use our disambiguation experiments with statistical query translation on a Greek-English cross language retrieval system using Google's n-grams. Results from our participation on the Ad-Hoc TEL track of CLEF 2009 are reported.

1. Introduction

Medas¹ is a Greek-English cross language retrieval system that aims to support Greek users in the medical domain, to overcome the language barrier. It contains two subsystems: a multilingual subsystem, for retrieving bilingual documents (a collection of scientific articles in medicine available in the Greek web) and a cross language subsystem, which provides only the interface to the MEDLINE database using the PubMed search engine. Medas contains a dictionary based translation module and uses the MeSH thesaurus for on-line reformulation of the queries. Preprocessing the queries includes language identification, tokenization, capital-to-lower letter conversion, stopword removal and stemming. Our participation in the Ad-Hoc multilingual retrieval track of CLEF this year is a first attempt to evaluate our system on a general-purpose (non-medical) set of queries.

A Cross Language Information Retrieval (CLIR) system needs an online and shallow translation system. This translation step tends to cause a reduction in cross language retrieval performance as compared to monolingual retrieval. Such an approach phases the barrier of the lexicon and has to overpass the polysemy problem in languages. Thus the translation is achieved in two phases: a translation and a disambiguation phase. For the translation of the initial query we shall use a bilingual term list and the disambiguation phase is based on the target language model, which is based on the statistical properties of n-grams.

For query translation a Greek-English bilingual term list is utilized as the main source of knowledge and Google's n-grams collection [1] is used for WSD to acquire the most appropriate translation.

The remainder of this paper is organized as follows. In the next section we briefly review the work on Greek-English CLIR. We continue in section 3 with the WSD using the web as a corpus. In section 4 we present our experimental framework and our experimental results. Conclusions are summarized in the final section.

2. Query translation and disambiguation

Automatic Machine Readable Dictionaries (MRDs) query translation, on its own, has been found to lead to a drop in effectiveness of 40-60 % of monolingual retrieval (Ballesteros and Croft [2], Hull and Grefenstette [3]). In our approach, query translation/disambiguation tasks are performed after a simple *stemming* process of source query terms to replace each term with its inflectional root, and to remove stop words. Then a term-by-term

¹ <http://www.medas.gr:8084/Medas/>

translation using dictionary-based method [4, 5, 6, 7], where each term or phrase in the query is replaced by a list of all possible translations, is completed.

Preprocessing Query includes language identification, tokenization, capital-to-lower letter conversion, stopword removal and stemming. This forces us to keep a stemmed copy of the dictionary. The dictionary contains single words and phrases –idioms, that is two or more words which translate to a single term and vice versa.

One of the main factors that limit the performance of retrieval is the limited coverage of the dictionary used in query translation. Ambiguity in translation of queries is one of the major causes for large drops in effectiveness below monolingual performance, for the dictionary-based method in CLIR.

Figure 1 presents the methodology that we used to perform the translation of the queries. Two sets of experiments were performed. In the first set we used as possible translations the translations that were derived from the dictionary whereas in the second we embodied in the candidate translations the synsets taken from WordNet [8].

In the following paragraphs of the current section we briefly describe the algorithms used for the word sense disambiguation, that is the language modelling and the maximum entropy model.

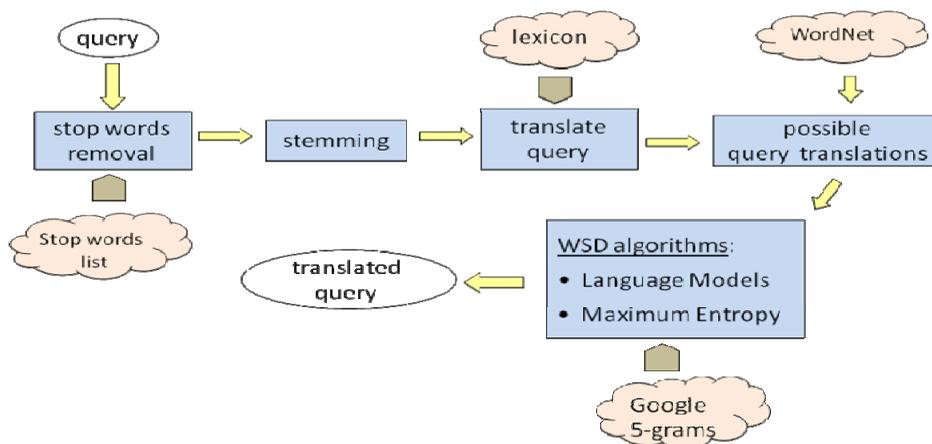


Figure 1. Overview of the WSD methodology

2.1 Language Modeling

The quality of n-gram language models depends directly on the size of the training texts. Thus the web has become a valuable resource in order to build better statistical language models.

By the term n-grams we define a sequence of n single words. In language models each word depends only on the context of k words. By the definition of the n-gram language models they have their own limitations as they can support dependency ranges up to (n - 1) tokens.

When estimating the probability of a translation candidate, it is impossible to use maximum likelihood estimates on the entire string, by using the chain rule

$$p(e) = p(e_1)p(e_2 | e_1)p(e_3 | e_1, e_2) \dots p(e_n | e_1, \dots, e_{n-1})$$

where e is a string in target language of length n, as one would run into serious data sparseness issues. Instead most language modeling approaches use a limited horizon of two words, called tri-gram models:

$$p(e) = p(e_1)p(e_2 | e_1)p(e_3 | e_1, e_2)\dots p(e_n | e_{n-2}, e_{n-1})$$

Despite the limited context that is taken into account they still form a good compromise between prediction quality, robustness and computational tractability.

2.2 Maximum Entropy Model

The second WSD method used is based on the calculation of the entropy of the translations with respect to the language model in the web [9]. Given a translation X of a question q , let's define w the sequence of n words that compose the translation $w=(w_1, \dots, w_n)$. A trigram chain is, therefore, defined as the set of trigrams T :

$$T = \{(w_1, w_2, w_3), (w_2, w_3, w_4), \dots, (w_{n-2}, w_{n-1}, w_n)\}$$

The general formulation of the information entropy is:

$$H(X) = -K \sum_{i=0}^n p(i) \log p(i) \quad (1)$$

where K is an arbitrary constant which depends on the problem, i is a fragment of message X of length n and, $p(i)$ is the probability of the i -th fragment. In the case of machine translation, the message is represented by the translation, and each fragment i corresponds to the i -th trigram of translation t_i . The probability of each trigram is calculated by means of web counts. Let's define the i -th trigram $t_i = (w_i, w_{i+1}, w_{i+2})$ and its root bigram $b_i = (w_i, w_{i+1})$. Let's name as $c(x)$ the function that returns the number of Google 5-grams that contain the text fragment x . According to [10], the probability $p(t_i)$ can be estimated as:

$$p(t_i) = \frac{c(t_i)}{c(b_i)} \quad (2)$$

If we substitute $p(i)$ with Formula 2 in Formula 1 and use a linear normalization factor as K , we obtain the formula that we used to calculate the entropy of a translation X :

$$H(X) = -\frac{1}{n} \sum_{i=0}^n \frac{c(t_i)}{c(b_i)} (c(t_i) - c(b_i)) \quad (3)$$

The selection of the best translation is made on the basis of the $H(X)$ calculated by means of Formula 4. Given M translations of q , we pick the translation m' such that:

$$m' = \arg \max_{m \in M} H(m) \quad (4)$$

3. The web as a corpus

Google released a collection of English n -gram data in August 2006. This data has been collected from the web pages and contains billions of n -gram types up to a maximum order of 5. The data are available from the Linguistics Data Consortium as a set of 6 DVDs. All together, the compressed n -gram data is about 24 GB in size, the uncompressed version takes more than 90 GB of space.

In all our runs we used only the titles of the topics. Thus we had very short queries of 2.5 words in average each, which made the disambiguation step difficult in certain cases. The average length of the translated queries using only our Greek-English lexicon, before disambiguation, was 7,6 words. We submitted 5 runs: two of these used a 2-gram language model for disambiguation, two used a 3-gram model and the final was used a maximum entropy model. Our results are summarized in figure 2 (mean values of precision at the 11 recall values). As a

conclusion we observed that the 2-gram model was best performed although in general our results were very poor compared to other participants.

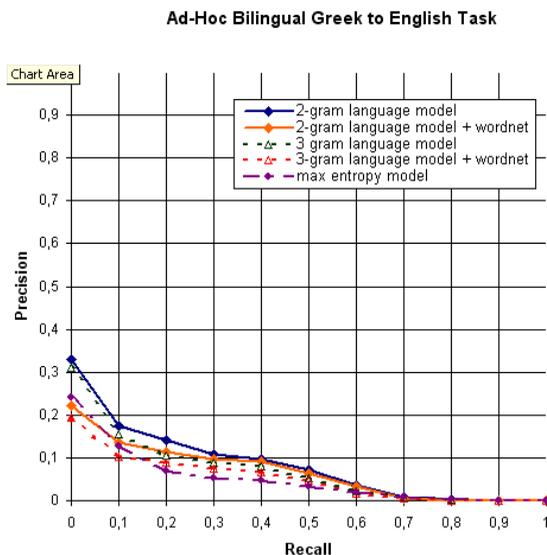


Figure 2. Mean interpolated precision at 11 recall values

4. References

1. LDC Catalog, Google n-grams, <http://www ldc upenn edu Catalog/CatalogEntry.jsp?catalogid=LDC2006T13>
2. Ballesteros, L. and Croft, W. B.: "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval". In proceedings of the 20th ACM SIGIR Conference, 1997, pp. 84-91.
3. Hull, D. and Grefenstette, G.: "Querying across languages. A Dictionary-based Approach to Multilingual Information Retrieval". In proceedings of the 19th ACM SIGIR Conference, 1996, pp. 49-57.
4. Sadat, F., Maeda, A., Yoshikawa, M. and Uemura, S.: "Integrating Dictionary-based and Statistical-based Approaches in Cross-Language Information Retrieval". IPSJ SIG Notes, 2000-DBS-121/2000-FI-58, 2000, pp. 61-68.
5. Adriani, M. 2000. Using statistical term similarity for sense disambiguation in cross-language information retrieval. *Information Retrieval*. 2, 69-80.
6. Gao, J., Nie, J. Y., Zhang, J., Xun, E., Zhou, M., and Huang, C. 2001b. Improving query translation for CLIR using statistical Models. In: *SIGIR '01*, New Orleans, Louisiana, pp. 96-104.
7. Christopher D. Manning, Hinrich Schütze, *Foundation of Statistical Natural Language Processing*. The MIT Press. pp229-262. 1999.
8. Fellbaum, C. (ed.): *Wordnet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press (1998)
9. Rosso, P., Buscaldi D., Iskra M. Web-based selection of optimal translations of short queries. In: *Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, num. 38, pp.49-52
10. Zhu, Xiaojin y Ronald Rosenfeld. 2001. Improving trigram language modeling with the World Wide Web. *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.