Word Sense Discrimination Using Context Vector Similarity

Atelach Alemu Argaw
Stockholm University/KTH
Stockholm, Sweden
atelach@dsv.su.se

Abstract

This paper presents the application of context vector similarity for the purpose of word sense discrimination during query translation. The random indexing vector space method is used to accumulate the context vectors. Pair wise similarity of the context vectors of ambiguous terms with that of anchor terms indicated the possible correct translation of a query term. Two retrieval experiments were conducted using the discriminated queries and weighted maximally expanded queries. The discriminated queries show a substantial increase in retrieval performance.

1 Introduction

In Cross Language Information Retrieval (CLIR) queries are posed in a language different from that of the document collection. Automatic translation of the queries to the language of the document collection is the most commonly used approach attributing to the fact that it is the least computationally expensive and the least resource intensive, when compared to the alternatives of document translation and Interlingua representation. One of the issues that need to be addressed during query translation is that of word sense disambiguation/discrimination where the task is to pick out the term with the correct sense in the current context among the possible translations given in the lexical resource utilized. Word sense disambiguation involves sense labeling and classifying or categorizing words into appropriate sense classes. Word sense discrimination (WSD) on the other hand, is not concerned about sense labeling, and aims to group the occurrences of a word into a number of classes by determining for any two occurrences whether they belong to the same sense or not. For Machine Readable Dictionary (MRD) based CLIR, we need to automatically determine the translation with the correct sense. In this paper we present an approach based on context vector similarity measures in a word space generated by random indexing.

2 Experiments

In information retrieval overall performance is affected by a number of factors, implicitly and explicitly. To try and determine the effect of all factors and tune parameters universally is a very complicated task. Therefore, our focus is in performing univariate sensitivity tests aimed at optimizing a specific single parameter while keeping the others fixed at reasonable values. In these experiments, our focus is to use context vector similarity in order to discriminate among potential translations, and determine the effect of such WSD in retrieval performance. Therefore, stemming of the Amharic terms is done manually in order to avoid the inclusion of wrong translations that might potentially misguide the discrimination process. Out of dictionary words are assumed to be named entities and are used as fuzzy match terms.

2.1 Data

We used the Cross Language Evaluation Forum (CLEF[1]) English collection, consisting of the LAT94 (56,472 articles) and GH95 (113005 articles) to generate the word space for sense discrimination as well as to perform the retrieval experiments. 50 Amharic queries from the CLEF 2004 were used for the discrimination and retrieval experiments.

[1] www.**clef**-campaign.org/

2.2 Preprocessing

For ease of use and compatibility purposes, the Amharic queries originally written using the Amharic script *fidel* were transliterated to an ASCII representation using a file conversion utility called g2[2]. Query terms that appear more than 15 times were discarded as they turned out to be non content bearing words and introduce noise when translated.

The query translation was done through term-lookup in an Amharic-English MRD (Aklilu, 1981) containing 15,000 entries and an online Amharic-English dictionary[3] with approximately 18,000[4] entries. Bigrams were given precedence for lookup over unigrams. The translated list of terms was further preprocessed by removing stop words and morphological analysis using the Porter stemming algorithm[5]. The document collection is also morphologically normalized using the same algorithm.

2.3 Word Sense Discrimination

Given an Amharic query with N terms $\{w_{a1}, w_{a2}, w_{a3}, \ldots, w_{an}\}$ and corresponding English translations $\{w_{a1e1}, w_{a1e2}, w_{a1e3}; w_{a2e1}; w_{a3e1}, w_{a3e2}, w_{a3e3}, w_{a1e4}, w_{a1e5}; \ldots; w_{ane1}, w_{ane2}\}$ we need to select the correct translation in the given context for the terms $\{w_{a1}, w_{a3}, \ldots, w_{an}\}$. The term w_{a2} is labeled as an 'anchor' term since it has a single term translation. Each possible translation is then paired with the closest anchor term, in this case $\{(w_{a1e1}, w_{a2e1}), (w_{a1e2}, w_{a2e1}), (w_{a1e3}\ w_{a2e1}) \ldots\}$. In the cases where there are two anchor terms at the same distance as a term that needs discrimination, precedence is given to the anchor occurring before the current term in the query, and in cases where there are no anchor terms, each unique pairing will be taken. We then calculate the similarity of the context vectors for each of the pairs of translated words, and from each group of translations for a single term, we pick the translation that has the highest similarity with the nearest anchor term. The pair with the highest similarity will be taken in the absence of anchors.

We used GSDM[6] (Guile Sparse Distributed Memory) and the morphologically normalized CLEF English document collection to generate the word space model using Random Indexing and train context vectors for syntagmatic word spaces (Sahlgren, 2006). We calculate the cosine similarity between the context vectors of each pair using the functions provided in GSDM.

2.4 Retrieval Experiments

The retrieval experiments were conducted using Lucene[7]. Two runs were designed for comparison purposes. Run1 used the discriminated queries and Run2 used a maximally expanded and weighted queries. In Run2, all translations for each word as given in the MRD are taken, but lesser weights were assigned to those terms with multiple translations. If a term has *n* possible translations, each of those translations are down scaled by assigning a weight of $1/n$ to each of the terms. The results are given in the table below.

	Relevant Total	**Relevant Retrieved**	**map**	**R-Prec**
Run1	375	163	11.91	10.76
Run2	375	129	5.79	4.83

Table 1: Retrieval Results

3 Concluding Remarks

The results obtained show that there is a substantial increase in retrieval performance when using the discriminated queries. The context vector similarity based discrimination works very well in picking out the correct translation. Further experiments using 150 more queries are currently underway to give the results more statistical significance and draw conclusions from.

Acknowledgments

The GSDM tool and the Guile functions were provided by Anders Holst and Magnus Sahlgren.

References

Amsalu Aklilu. 1981. *Amharic English Dictionary*, volume 1. Mega Publishing Enterprise, Addis Ababa, Ethiopia.

Magnus Sahlgren. 2006. *The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Doctoral thesis, Stockholm University.

[2] g2 was made available to us through Daniel Yacob of the Ge'ez Frontier Foundation (http://www.ethiopic.org/)

[3] http://www.amharicdictionary.com/

[4] Personal correspondence

[5] http://tartarus.org/~martin/PorterStemmer/index.html

[6] GSDM is an open source C-library for Guile, designed specifically for the Random Indexing methodology, written by Anders Holst at the Swedish Institute for Computer Science.

[7] Apache Lucene is an open source high-performance, fullfeatured text search engine library written in Java http://lucene.apache.org/java/docs/index.html