Using a Cross-Language Information Retrieval System based on OHSUMED to Evaluate the Moses and KantanMT Statistical Machine Translation Systems

Nikolaos Katris[1], Richard Sutcliffe[2], Theodore Kalamboukis[3]

[1]Department of CSIS, University of Limerick, Limerick, Ireland
[2]School of CSEE, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK
[3]Department of Informatics, Athens University of Economics and Business, 76 Patission Str., Athens 10434, Greece
E-mail: nikos.katris@live.com, rsutcl@essex.ac.uk, tzk@aueb.gr

Abstract

The objective of this paper was to evaluate the performance of two statistical machine translation (SMT) systems within a cross-language information retrieval (CLIR) architecture and examine if there is a correlation between translation quality and CLIR performance. The SMT systems were KantanMT, a cloud-based machine translation (MT) platform, and Moses, an open-source MT application. First we trained both systems using the same language resources: the EMEA corpus for the translation model and language model and the QTLP corpus for tuning. Then we translated the 63 queries of the OHSUMED test collection from Greek into English using both MT systems. Next, we ran the queries on the document collection using Apache Solr to get a list of the top ten matches. The results were compared to the OHSUMED gold standard. KantanMT achieved higher average precision and F-measure than Moses, while both systems produced the same recall score. We also calculated the BLEU score for each system using the ECDC corpus. Moses achieved a higher BLEU score than KantanMT. Finally, we also tested the IR performance of the original English queries. This work overall showed that CLIR performance can be better even when BLEU score is worse.

Keywords: cross-language information retrieval, statistical machine translation, Moses, KantanMT, Apache Solr, EMEA, QTLP, ECDC, OHSUMED

1. Objective

Information retrieval (IR) (Grefenstette 1998) has nowadays a prominent place in our everyday lives. People of diverse backgrounds from all over the globe use search engines to look for information for nearly every imaginable human need. The word "Google" has been introduced as a verb in many natural languages, referring to the use of the famous search engine to obtain information from the World Wide Web[1].

Nevertheless, the Web is not the only field that requires the deployment of information retrieval systems. Searching for an email in our inbox or looking for a document within the large corporate intranet of a company containing specific keywords also represent examples of IR tasks. However, it was the rapid expansion of the World Wide Web that brought to the forefront a major challenge that IR systems needed to overcome. As the Web began to grow, the information posted started to vary in regard to language, with the amount of non-English content constantly increasing (Peters et al 2012).

As a result, much information is being disregarded because it is available in a less popular language, i.e. not in English. And while most content on the Web is still available in English (over 50%[2]), the assumption that most users of the Internet (or most users of any IR system within an organization) have a good knowledge of the English language is not always true. In fact, foreign language skills vary significantly depending on various factors, such as country of residence, educational background, etc. And, of course, there are those who are able to understand a foreign language, but are unable to adequately write a query in that language.

It has thus become apparent that monolingual information retrieval is not able to meet the language requirements of a multilingual world. Cross-language information retrieval (CLIR) is trying to bridge the gap when it comes to searching for something in one language and retrieving relevant documents in another. Succinctly said, a CLIR system deploys the same methods as a monolingual IR system, but it also uses a translation module to convert the queries or the documents from one language to another (Nie 2010).

The objective of this paper is to evaluate the performance of two statistical machine translation (MT) platforms within the same information retrieval scenario and determine whether there is a relationship between the quality of the machine translation and the CLIR performance. Moses [3] and KantanMT [4] were the MT systems used to translate a series of queries from Greek into English and the English queries were used to search and retrieve relevant documents from a document collection using Apache Solr.

2. Greek Language in CLIR

Since we focus on the Greek-English language combination, it would be useful to give a short presentation of the issues that Greek poses during CLIR. The Greek language belongs to the Indo-European family of languages. It is the official language of Greece and Cyprus, it is one of the official languages of the European Union and is spoken by approximately 13 million people. It also has the longest history of any Indo-European language, with over thirty centuries of written records.

Compared to English, the computational processing of Modern Greek is a highly difficult task. Its inflectional and conjugational characteristics, with a plethora of endings,

[1] https://www.google.com/

[2] http://w3techs.com/technologies/overview/content_language/all

[3] http://www.statmt.org/moses/

[4] https://kantanmt.com/

suffixes and prefixes as well as the accents, which also change their position depending on the inflection, result in *4-7 word forms for a noun and up to 250 word forms for a verb* (Karanikas et al 2000). These endings are thus essential in defining the grammatical and syntactic role of the individual tokens, especially considering that Greek follows less strict rules regarding the position of the syntactic elements within a sentence. The level of difficulty of text processing increases further if we take into account any archaic word forms or expressions that are commonly used in Modern Greek. On the other hand, the same inflectional characteristics contribute to the significant reduction of ambiguity that is for example common in English.

Greek is not one of the most researched languages with regard to CLIR although its aforementioned derivational and inflectional characteristics are the main source of issues regarding any NLP-related task involving Greek. Dictionary-based approaches to CLIR involving Greek face numerous problems. Due to the morphological complexity of the language, machine-readable dictionaries have limited coverage. Coverage can be improved using stemming techniques, although they tend to increase the level of uncertainty since more words with different meanings are conflated into the same stem. Therefore, there are two levels of uncertainty which a CLIR system involving Greek has to deal with: one caused by the stemming process and one due to the natural ambiguity of the language (words with more than one possible translation) (Kotsonis et al 2008).

MedAS (Medical Assistance System) is a Greek-English cross language retrieval system that aims to help Greek users who work in the medical domain to overcome the language barrier. According to Katsiouli and Kalamboukis (2009) MedAS "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[5] database using the PubMed[6] search engine". MedAS also uses a dictionary-based translation module together with the MeSH[7] thesaurus for online reformulation of the queries.

Unfortunately, no information was found regarding the utilization of an SMT approach in Greek CLIR scenarios, which is one the reasons for including Greek in the experiments of this paper.

3. Experimentation Outline

There is an abundance of methods and approaches to the implementation of CLIR. The most common ones involve some sort of translation of either the query or the documents. Because of its flexibility and its lower computational requirements the experiments were conducted using the query translation method (Peters et al 2012). The translated version (of either the query or the documents) is then used for indexing and the indexed representation is used for matching and retrieving the relevant documents. Each method has, of course, its own advantages and disadvantages. A set of queries was translated from Greek into English using KantanMT and Moses and used as input to Apache Solr, the IR tool, for retrieval of relevant documents from an English document collection. The results from the two sets of translated queries (one set of translated queries for each MT system) were then compared to the gold standard, i.e. the documents which are truly relevant, in order to calculate the precision, recall and F-measure values for each system. Moreover, we examined if there is a correlation between translation quality and IR performance using the BLEU metric as the quality score. Finally, we also calculated the precision, recall and F-measure values produced when feeding the IR system with the corresponding English human-produced queries.

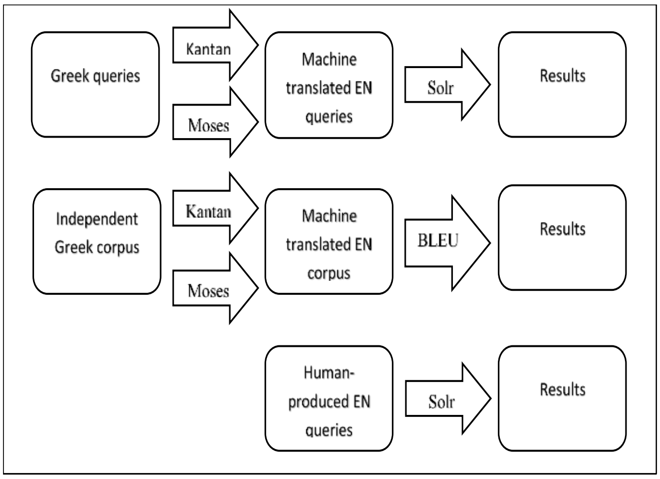

Figure 1: The three experiments

Figure 1 illustrates the three experiments that we ran. The first one is the CLIR experiment, which compares the information retrieval performance of each set of machine translated queries (one set translated by KantanMT and one set translated by Moses). The second experiment determined the quality that each system offered using the BLEU score on an independent corpus. The third experiment evaluates the performance of the original English queries in the context of a monolingual IR experiment.

Figure 2 gives an overview of the whole CLIR system architecture as it was structured for the purposes of this paper.

3.1 Language resources

Besides the three basic tools (Moses, KantanMT, Apache Solr) that were used to run the experiments, a series of other linguistic resources as well as minor tools were also needed for various purposes. If we divide the CLIR experiment into two modules, one about the machine translation and one about the information retrieval ("loose coupling" as described by Peters et al, 2012), then the following data were required in order to use each one:

- Machine translation module
    - Parallel bilingual Greek-English corpus

[5] https://www.nlm.nih.gov/pubs/factsheets/medline.html
[6] http://www.ncbi.nlm.nih.gov/pubmed
[7] https://www.nlm.nih.gov/mesh/

for MT training
    - Parallel bilingual Greek-English corpus for MT tuning
    - Monolingual English corpus for language modelling
    - Bilingual Greek-English corpus for BLEU score calculation
- Information retrieval module
    - Set of Greek queries
    - English document collection
    - Gold standard (the correct answers, i.e. relevant documents for to each query).

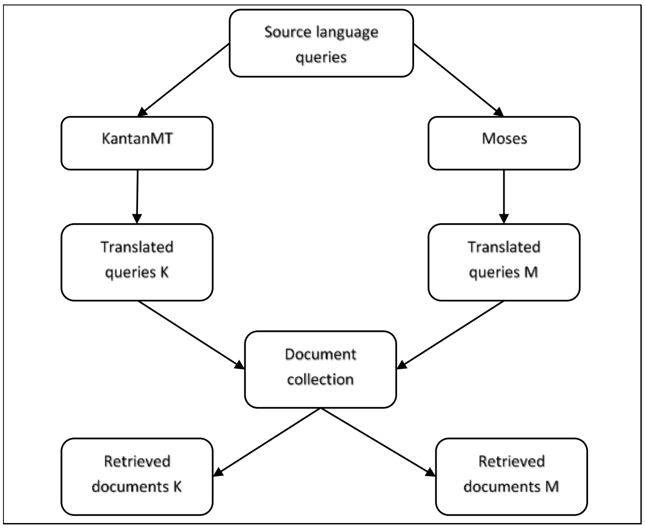

Figure 2: CLIR architecture overview

3.1.1. Test collection

The document collection used was the OHSUMED database (Hersh et al 1994). The OHSUMED test collection is a subset of MEDLINE, the online medical information database that contains 233,445 references consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991). Each document contains descriptive fields, namely sequential identifier (.I), MEDLINE identifier or document id (.U), human-assigned MeSH (Medical Subject Headings) indexing terms (.M), title (.T), publication type (.P), abstract (.W), author (.A) and source (.S). Only the abstract field was used for indexing of the files.

Apart from the documents, OHSUMED contains two other very important resources: a set of queries and their answers, i.e. the relevant documents to the queries. The queries were developed for the purposes of experiments conducted by Hersh et al (1994) and originally they were 106 in total. However, for this paper we used the OHSUMED database as provided by Text REtrieval Conference (TREC) and, since for the TREC tasks only a subset of 63 queries were used, the same path was also followed in the experiments.

The relevance assessments were provided by physicians who were clinically active and were current fellows in general medicine or medical informatics or senior medical residents (Hersh et al 1994). They were asked to determine the relevance on a three-point scale: definitely relevant, possibly relevant and not relevant. Their judgements were used as the gold standard for this experiment.

Since the queries contained in OHSUMED are in the English language, we needed an objective approach to how we could use this test collection in CLIR involving the Greek language. Since no CLIR test collection with Greek queries was available, the obvious approach was to translate the queries into Greek (using a human translator) and then translate them back into English using the two SMT applications. That way, we assumed that the Greek queries, if translated "perfectly" into English, would return the relevant documents as defined for the English queries.

A Greek version of the 63 original OHSUMED queries was used, offered from (Kotsonis et al 2008). These queries were translated by an independent Greek medical doctor and ensured that the experiments were unbiased. The queries were formulated in a way a typical Greek medical doctor would search for the corresponding information. This was a very important factor for properly conducted experiments.

3.1.2. Training Corpora

In order to translate the queries from Greek into English with KantanMT and Moses, we needed to train the systems using a series of corpora. Because the test collection –and therefore the queries– was from the medical domain, the corpora used for training the SMT systems should also come from the same domain. We used three kinds of corpora for MT training:

1. A bilingual parallel Greek-English corpus for the translation model
2. A bilingual parallel Greek-English corpus for tuning
3. A monolingual English corpus for the language model
4. A bilingual parallel Greek-English corpus for testing (calculating the BLEU score)

The corpus that was used for the creation of the translation model for each of the MT systems was the EMEA corpus (Tiedemann 2009). The EMEA corpus is a parallel bilingual corpus provided by the European Medicines Agency (EMA). It consists of 1,073,225 sentence pairs and 24,670,000 words and it is available in 22 languages, however for the purposes of these experiments we used the Greek-English combination.

The target language TXT file of the same corpus was used for the purposes of building the language model in both KantanMT and Moses.

For the tuning process a separate parallel bilingual corpus was needed. For this purpose we used the QTLP English-Greek Corpus for the medical domain, which was downloaded from the META-SHARE repository[8]. It contains automatically detected pairs of parallel documents that were acquired from the web during 2013-14 using the ILSP Focused Crawler[9], an open source tool that was enhanced in the context of QTLP by researchers of the Institute for Language and Speech Processing, Athens. After the processing and export of the documents, the aligned pairs of sentences were extracted from pairs of parallel documents using a sentence alignment web service hosted by DCU[10]. The corpus has 62,452 pairs of aligned

[8] http://qt21.metashare.ilsp.gr/repository/browse/qtlp-english-greek-corpus-for-the-medical-domain/665f3832a93211e3b7d800155dbc020119068d5402fc4d3bb497aa9dcd7a4892/

[9] ILSP-FC, http://nlp.ilsp.gr/redmine/projects/ilsp-fc

[10] http://srv-cngl.computing.dcu.ie/panacea-soaplab2-

sentences of 1,234,556 English tokens and 1,275,151 Greek ones.

Finally, we used the ECDC Greek-English Translation Memory subcorpus as the testing corpus, which was used for the calculation of the BLEU score. The ECDC corpus is a translation memory provided by European Union (EU) agency "European Centre for Disease Prevention and Control" (ECDC). It contains 2,469 sentences of 44,315 words in total.

4. Experiments

Three experiments were run for the purposes of this paper:

1. A CLIR experiment, which compares the information retrieval performance of each set of machine translated queries (one set translated by KantanMT and one set translated by Moses) using the OHSUMED test collection.
2. A BLEU score calculation experiment for both MT systems using the ECDC corpus.
3. A monolingual IR experiment using the original English queries from the OHSUMED test collection.

4.1 Experiment 1

For the CLIR experiment, we began by training KantanMT and Moses using the same training data:

1. For the translation model we used the EMEA corpus
2. For the language model we used the EMEA English sentences
3. For tuning we used the QTLP corpus

After the successful completion of both training processes, we took the Greek queries that were previously translated from English by an independent Greek medical doctor (Kotsonis et al 2008), and translated them using KantanMT and Moses.

Having two sets of 63 machine translated English queries, we moved on to the information retrieval part. The 233,445 documents of the OHSUMED test collection were loaded and indexed in Apache Solr[11].

Then, each translated query (63 from KantanMT and 63 from Moses) was run in order to retrieve relevant documents. Only the top 10 results (n=10) were taken into account for the evaluation of the retrieval of each query.

We calculated average precision, recall and F-measure for all 63 queries. KantanMT produced a higher average F-measure compared to Moses (0.07 vs 0.05 as shown in Table 1). The average precision was also higher for KantanMT (0.12 vs 0.10), while recall was the same for both tools (0.06). Nevertheless, the MT systems didn't always produce corresponding results for each individual query. In some queries Moses performed better than KantanMT. The average precision was also higher than the average recall for both tools. KantanMT had an average precision of 0.12 and an average recall of 0.06, while Moses scored a 0.10 precision and 0.06 recall (Table 1).

KantanMT			Moses		
P	R	F	P	R	F
0.12	0.06	0.07	0.10	0.06	0.05

Table 1: Experiment 1 results

4.2 Experiment 2

The aim of the second experiment was to evaluate the quality of the translated output from both KantanMT and Moses using an independent corpus as a test set.

The ECDC corpus was used as a test set for calculating the BLEU score (Papineni et al. 2001) of each MT system. The Greek part of the corpus was translated using both KantanMT and Moses and the outputs from each system were compared to the reference translations contained in the original English file.

Moses achieved a 17.72 BLEU score vs. the 11.74 of KantanMT. These results confirm the theory that there is not necessarily a correlation between translation quality and IR performance, because KantanMT scored more highly in Experiment 1.

	BLEU score
KantanMT	11.74
Moses	17.72

Table 2: Experiment 2 results

4.3 Experiment 3

The aim of the third experiment was to compare the CLIR performance using KantanMT and Moses with the performance of the original English queries. This would give us a better understanding of the performance of the two CLIR scenarios (one with KantanMT and one with Moses) against the monolingual scenario using human-produced English queries.

The 63 original English queries were directly entered into Apache Solr and executed in order to retrieve relevant documents from the 233,445 indexed documents of OHSUMED test collection.

The results showed that, once again, average precision was higher than the average recall (0.22 vs 0.12 in Table 3). As expected, all three scores were almost double compared to the ones the machine translated queries were able to achieve. As shown in Table 3, the F-measure for the original English queries is 0.13, which is around double the values for KantanMT (0.07) and Moses (0.05) in Table 1.

Original English Queries		
P	R	F
0.22	0.12	0.13

Table 3: Experiment 3 results

Nevertheless, it is interesting to note that for certain queries the machine translated queries achieved higher scores. For example, the eleventh query (OHSU11) had an F-measure of 0.16 for KantanMT and 0.19 for Moses. However, the original English query managed to yield an F-measure of only 0.08, which is half the performance of the machine translated queries.

Another interesting aspect that concerns the monolingual IR part of the experiment is that even though the performance is higher, the numbers are still pretty low and there are several cases where no relevant document was retrieved whatsoever. There can be several reasons for this, such as the settings of Apache Solr used for indexing and matching and the number of the top retrieved documents that were taken into account when calculating precision,

axis/#panacea.hunalign_rowhunalign

[11] http://lucene.apache.org/solr/

recall and F-measure (n=10).

Moreover, for the purposes of our experiment we only used the abstract of each document. However, it is possible that if we had also included the title of each document, the system could possibly return a higher number of relevant documents. The same applies, of course, to the CLIR scenario and may also explain the low numbers produced when using the queries translated by KantanMT and Moses.

5. Conclusion

The aim of this paper was to evaluate KantanMT and Moses as two statistical machine translation applications within a cross-language information retrieval architecture. KantanMT proved to be slightly better than Moses. Moreover, our experiments showed that the use of SMT in CLIR produces quite good results compared to a monolingual retrieval scenario.

The key result of this research, however, is that the BLEU score of an SMT system does not necessarily correlate with the results it generates in a CLIR system. In Experiment 1 (CLIR) KantanMT scored more highly than Moses, while in Experiment 2 (BLEU evaluation) Moses scored more highly than KantanMT.

Cross-language information retrieval using full SMT systems is not a highly researched field, particularly involving the Greek language. However, as SMT grows more and more popular we hope that its implementation in CLIR scenarios will also be further examined.

There are many suggestions for further experimentations and approaches to using SMT in a CLIR system that could enlighten the many facets of such an architecture. Numerous other settings could have been applied to the IR system in order to determine if there is any improvement in the retrieval. Another interesting aspect would be to check the performance of the machine translated queries depending on their length. Also, more research needs to be conducted regarding the presence of untranslated terms in the queries and how they affect the performance. Finally, it would be interesting to experiment with better maintained test collections and collections from other domains. Of course, these are only some of the many aspects that could be examined.

6. Acknowledgements

We are indebted to Stelios Piperidis of ILSP for his help with this work.

7. Bibliographical References

Grefenstette, G. (1998). Problems and Techniques of Cross Language Information Retrieval. Information Retrieval, (May), pp. 523–524.

Hersh, W., Buckley, C., Leone, T. J., Hickman, D. (1994). Ohsumed: an interactive retrieval evaluation and new large text collection for research. In Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, pp. 192–201.

Karanikas, H., Tjortjis, C., Theodoroulidis, B. (2000). An Approach to Text Mining using Information Extraction. Centre for Research in Information Management, Department of Computation.

Katsiouli, P., Kalamboukis, T. (2009). An Evaluation of Greek-English Cross Language Retrieval within the CLEF Ad-Hoc Bilingual Task. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2009 Workshop

Kotsonis, E., Kalamboukis, T. Z., Gkanogiannis, A., Eliakis, S. (2008). Greek-English cross language retrieval of medical information. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4987 LNCS, pp. 109–117.

Nie, J. W. (2010). Cross-Language Information Retrieval, in Cross-Language Information Retrieval, Synthesis Lectures in Human Language Technologies, Morgan & Claypool.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation, Technical report RC22176(W0109-022), IBM Research Report.

Peters, C., Braschler, M., Clough, P. (2012). Multilingual Information Retrieval - From Research To Practice, Springer.

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), John Benjamins, Amsterdam / Philadelphia, pp. 237-248.