Caption vs. Query Translation for Cross-Language Image Retrieval

Paul Clough
Department of Information Studies, University of Sheffield, Sheffield, UK.
p.d.clough@sheffield.ac.uk

Abstract

For many cross-language retrieval tasks, the predominant approach is to translate the query into the language of the document collection (target language). This often gives results as good as, if not better, than translating the document collection into the query language (source language). In this paper, I evaluate query versus document translation for the ImageCLEF 2004 bilingual ad hoc retrieval task. Image retrieval is achieved through matching textual queries to associated image captions for the following languages: French, German, Spanish and Italian. Captions and queries are translated using commercially and publicly available resources: `Systran Professional` and `Babelfish`, and retrieval performed using the KL-divergence language model provided by the `Lemur` toolkit. On average we find query translation to outperform document translation, but this varies across language and query.

1 Introduction

A great deal of research is currently underway in the field of Cross Language Information Retrieval (CLIR) where documents written in one language are retrieved by a query written in another (see, e.g. [1] and [2]). One can consider CLIR as basically a combination of machine translation (MT) and traditional monolingual information retrieval (IR). Most CLIR research has focused on locating and exploiting translation resources with which the users search requests or target documents (or both) are translated into the same language [3]. Monolingual approaches can then used for retrieval.

There are at least three methods for translation: (1) bilingual dictionaries, (2) extracting word/phrase equivalents from parallel or comparable corpora, and (3) using a Machine Translation (MT) system. Advantages and disadvantages exist for each approach, including the degree of linguistic resources and knowledge required for translation. Dictionary-based methods dominate query translation, but these often require extensive language processing to deal with issues such as lexical ambiguity, morphological variation, orthography, tokenisation and compound word splitting. MT approaches have proven to be popular in recent years due to the availability of on-line MT systems which can be exploited for query translation [1]. The MT system can often be treated as a "black box" where a single translation is provided from the input query. This can be a disadvantage for query translation where short, ungrammatical queries can be mistranslated due to limited context. However, an advantage of MT methods is that little or no further linguistic processing or resources are necessary to produce usable CLIR systems (see, e.g. [4]).

In document translation, the entire collection is first translated prior to searching. Previous research by Oard [5] showed that for German–English TREC-6 data, MT-based query translation out-performed various dictionary-based methods, and document translation out-performed MT query translation, especially for longer queries. McCarley [6] showed that for French–English

[1] See, for example, the large number of submissions in CLEF 2003 which utilised on-line MT systems

TREC-6 and TREC-7 data and using a statistical MT method, retrieval effectiveness was influenced by the direction of translation (French–English performed better than English–French for query and document translation). Fujii and Ishikawa [7] presented a two-stage method where initial retrieval was first performed using query translation, then the top 1000 documents translated into the query language using MT, finally documents re-ranked based on a translation score. This method was shown to outperform query translation alone and be well suited to large collections.

Advantages of document translation include: (1) no query translation is required at run-time, and (2) no further translation is required when presenting the results to the user. However, a major disadvantage is that translation of large collections is expensive both in time and resources. For example, Oard [5] spent ten machine-months translating the SDA/NZZ German collection (251,840 newswire articles).

One area of CLIR research which has received less attention is image retrieval. In collections such as historic or stock-photographic archives, medical case notes and art/history collections, images are accompanied by some kind of text (e.g. meta-data or captions) semantically related to the image [2] [12]. Images can then be retrieved using standard text-based IR methods. For those organisations managing image repositories in which text is associated with images (e.g. on-line art galleries), one way to exploit these is by enabling multilingual access to them. Given that image captions are typically much smaller than standard test-collection documents, it is feasible to perform document translation, even on larger image collections.

In this paper I compare query and document translation for a cross-language image retrieval task based on the St. Andrews collection of historic photographs and topics from the ImageCLEF 2004 ad hoc retrieval task. This paper divides into the following: section 2 describes the experimental setup, section 3 describes the results, and section 4 concludes this experiment.

2 Experimental Setup

In this section I describe the experimental setup including the ImageCLEF test collection, the `Lemur` retrieval system, the `Systran` and `Babelfish` Machine Translation tools, and additional language processing used for cross-language retrieval.

2.1 The ImageCLEF 2004 Ad Hoc Test Collection

The ImageCLEF 2004 ad hoc test collection consists of a document collection, a set of user needs expressed in both natural language and with an exemplar image, and for each user need a set of relevance judgements [8]. Topics and relevance judgements are provided for an ad hoc retrieval task which is this: given a multilingual statement describing a user need, find as many relevant images as possible from the document collection. This retrieval task simulates when a user is able to express their need in natural language, but requires a visual document to fulfil their search request.

The document collection consists of 28,133 images from the St Andrews Library photographic collection[2] and all images have an accompanying textual description consisting of 8 distinct fields (see, e.g. Figure 1). These fields can be used individually or collectively to facilitate image retrieval. The 28,133 captions consist of 44,085 terms and 1,348,474 word occurrences; the maximum caption length is 316 words, but on average 48 words in length. All captions are written in British English, although the language also contains colloquial expressions. Approximately 81% of captions contain text in all fields, the rest generally without the description field. In most cases the image description is a grammatical sentence of around 15 words. The majority of images (82%) are in black and white, although colour images are also present in the collection.

The 2004 test collection consists of 25 queries (topics) designed to simulate a range of realistic search requests to a cross language image retrieval system. English versions of the topics consist of a title (a short sentence or phrase describing the search request in a few words), and a narrative (a description of what constitutes a relevant or non-relevant image for that search request). The

[2] http://specialcollections.st-and.ac.uk/photcol.htm (site visited: 05/08/2004).

Title: Old Tom Morris, golfer, St Andrews.
Short title: Old Tom Morris, golfer.
Location: Fife, Scotland
Description: Portrait of bearded elderly man in tweed jacket, waistcoat with watch chain and flat cap, hand in pockets; painted backdrop.
Date: ca.1900
Photographer: John Fairweather
Categories: `[golf - general]`, `[identified male]`, `[St. Andrews, Portraits]`,`[Collection - G M Cowie]`
Notes: GMC-F202 pc/BIOG: Tom Morris (1821-1908) Golf ball and clubmaker before turning professional, later Custodian of the Links, St Andrews; golfer and four times winner of the Open Championship; father of Young Tom Morris (1851-1875).DETAIL: Studio portrait.

Figure 1: An example image and caption from the ImageCLEF collection

titles of each topic have been translated into 12 languages: Spanish, Italian, German, French, Dutch, Danish, Swedish, Finnish, Chinese, Japanese, Russian and Arabic by native speakers. The test collection also consists of a set of relevance judgements for each topic based primarily on the image, but also assisted by the image caption.

2.2 The Lemur Retrieval System

In the Lemur implementation of language modelling for IR, documents and queries are viewed as observations from generative unigram language models (see, e.g. [9] for more information). Queries and documents are represented as estimated language models with word probabilities derived from the documents, queries and the collection as a whole. The estimated query and document language models ($\hat{\theta}_Q$ and $\hat{\theta}_D$ respectively) are compared and ranked using the KL-divergence measure, an approach which can be likened to the vector-space model of retrieval where queries and documents are represented by vectors rather than language models.

In these experiments, the KL-divergence language model is used with the absolute discounting method of smoothing ($\Delta = 0.7$). Lemur offers query expansion by supplementing the initial query with collection-specific terms obtained from a feedback model. In these experiments, a two-component mixture model is used to estimate word probabilities in the feedback model. Default parameter values are used for the feedback model: $\alpha = \beta = 0.5$, with 20 terms selected from the top 10 documents retrieved from the initial query (pseudo relevance feedback or PRF), with one feedback iteration.

2.3 Translation Resources

In these experiments, translation is performed using the Systran and Babelfish machine translation (MT) resources. The original English captions were translated into German, French, Italian and Spanish using Systran Professional Premium 3.0 which took about 2 hours for each language pair[3]. For query translation Babelfish from Alta Vista was used. This free on-line resource is powered by Systran thereby allowing comparison between query and document translation with the same resource.

Like any form of translation method, MT can result in erroneous queries because of difficulties encountered during translation including: short queries resulting in little if none syntactic structure to exploit, errors in the original cross language text (e.g. spelling mistakes or incorrect use of diacritics), lack of coverage by the translation lexicon, incorrect translation of phrases, mis-translation of proper names, and incorrect translation of ambiguous words (e.g. selecting the wrong sense of a noun or verb). The effect of translation errors on retrieval performance for ImageCLEF 2003 topics is discussed in [10]. For more information on Systran, see e.g. [11].

[3] Captions were translated by Jianqiang Wang and Doug Oard at Maryland University.

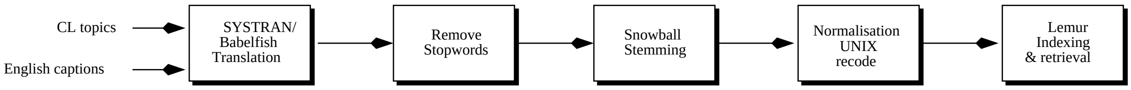
CL topics
English captions
SYSTRAN/ Babelfish Translation
Remove Stopwords
Snowball Stemming
Normalisation UNIX recode
Lemur Indexing & retrieval

Figure 2: The cross-language retrieval process

2.4 The Retrieval Process

Given multilingual captions and queries, we compare query versus document translation using the cross-language retrieval process shown in Figure 2. Given multilingual topics to translate into English, or English captions to translate into French, German, Italian or Spanish, we first translate the texts using the MT system. Next stopwords are removed using stopword lists provided with the `Snowball` stemmer[4]. To improve recall, we then apply stemming using `Snowball` and remove diacritics using the UNIX `recode` tool. To perform this, we recode the character set from latin1 to HTML (e.g. German topic 7 "Außenansichten von Tempeln in Ägypten" is transformed into "Au&szlig;enansichten von Tempeln in &Auml;gypten", and then the HTML characters are replaced by their original ascii characters (e.g. "Au*ss*enansichten von Tempeln in *A*gypten"). Finally, all characters are converted to lower case. So, for example, "Außenansichten von Tempeln in Ägypten" is reduced to "aussenansicht tempeln agypt". The resultant captions are indexed with `Lemur` and retrieved using the KL-divergence language model. All fields from the image caption are used during indexing.

We perform two retrieval experiments, one based on query translation and the other based on document translation. In both experiments we also perform retrieval with and without feedback and compare retrieval effectiveness for individual topics. Runs are distinguished as follows: de_de represents German document translation (i.e. German topics retrieved from the document collection translated automatically into German); de_en represents German query translation (i.e. German topics translated automatically into English and retrieved from the English document collection). Languages are identified as follows: fr = French, de = German, es = Spanish, en = English and it = Italian. Runs with an additional "fb" represent retrieval with pseudo relevance feedback.

2.5 Evaluation Measures

Uninterpolated Mean Average Precision (MAP), recall and a normalised precision at 100 measure $P_{norm100}$ are used to measure retrieval effectiveness. We use $P_{norm100}$ to compare runs with and without feedback because this measure is not influenced by changes in the rank position of relevant documents; only by the number of relevant which occur in the top 100 (from previous experiments we assume that users are willing to view at least 100 images [12]). $P_{norm100}$ normalises precision at 100 (P_{100}) with respect to the number of relevant documents for each query[5]. This measures the proportion of relevant documents retrieved in the top 100 (i.e. a recall measure), rather than the proportion of the top 100 which are relevant. Given a P_{100} score and a set of relevance judgements for a query Φ (the size given by $\mid \Phi \mid$), the normalised precision at 100 score $P_{norm100}$ is given by:

$$P_{norm100} = \frac{P_{100} \times 100}{min(100, \mid \Phi \mid)} \tag{1}$$

The normalised precision score ranges from 0 indicating no relevant in the top 100, to 1 which indicates either all relevant are in the top 100 (if $\mid \Phi \mid \leq 100$) or that all top 100 documents are relevant (if $\mid \Phi \mid > 100$). We define $P_{norm100} = 1$ as a *good* topic (further retrieval unrequired) and $P_{norm100} = 0$ as a *bad* topic (relevance feedback will be unsuccessful unless the user is willing to go beyond the top 100).

[4] http://snowball.tartarus.org/ (site visited: 04/08/2004)

[5] We explain this further and justify this measure in [12]

Table 1: A summary of monolingual retrieval effectiveness (i.e. caption translation)

	MAP	%English MAP	Recall	Avg $P_{norm100}$	Topics good	Topics bad
en_en	0.6185	-	0.9566	0.4388	7	0
en_en_fb	0.5829	-	0.9614	0.4016	5	1
de_de	0.3019	48.8%	0.7407	0.2870	5	2
de_de_fb	0.2769	44.8%	0.6791	0.2769	6	6
fr_fr	0.4328	70.0%	0.7817	0.3338	3	2
fr_fr_fb	0.4365	70.6%	0.8914	0.3432	6	2
it_it	0.3947	63.8%	0.7250	0.3072	4	1
it_it_fb	0.4355	70.4%	0.8552	0.3276	7	2
es_es	0.4836	78.2%	0.9469	0.3678	5	1
es_es_fb	0.4365	70.6%	0.8372	0.3634	6	2

3 Results

Results for all 18 runs are shown in the following tables: Table 1 summarises retrieval effectiveness for document translation, and Table 2 for query translation. On average across all multilingual runs, query translation outperforms document translation based on MAP: 0.4742 (75% of highest English MAP) vs. 0.3998 (65% of highest English MAP) respectively. This varies across language, however, where document translation for Spanish (without feedback) is higher than results for query translation (a MAP of 0.4836 vs. 0.4654 respectively). However, differences between results without feedback between document and query translation are statistically significant only for Italian and German (using the Wilcoxon test with $p < 0.05$).

Query translation is more successful because the translation pair X→English is typically better than English→X[6]. English→German performs worst and upon inspection we find that most errors are due to English words being incorrectly combined to form German compound terms. For example, the phrase “Falls of Cruachan Station above Loch Awe” is translated into “Fälle der StationCruachan über Lochawe”. In this example, “Cruachan Station” and “Loch Awe” are combined rather than left as proper names. We also find determiners and conjuctions are wrongly combined, e.g. “below embankment” translates to “unterDamm” and “lining banks” to “dieBänke”. Part of the problem is caused by captions texts being“dirty” and ungrammatical and could be improved by cleaning up the English texts prior to translation. This is less problematic for query translation resulting in higher retrieval effectiveness. Document translation is more successful for other languages because Spanish, French and Italian are less compound rich than German making X→English translation better.

In general results show that feedback reduces performance after the initial retrieval. Based on $P_{norm100}$, all results for query translation are lower with feedback. Results are similar for document translation except French and Italian for which results are higher (differences are not statistically significant). The most likely reasons for this are: (1) non-optimal parameter settings for the feedback model, and (2) few relevant documents in the top 10 being used for relevance feedback. In the remainder of this discussion we focus on results without feedback.

Figures 3 to 6 show average precision results for individual queries for both document and query translation as stacked bar graphs. Although in general document translation outperforms query translation for all languages except Spanish, it is interesting to observe that this is not the case for all queries. For retrieval in German, 9 queries perform better with document translation, 10 for French, 11 for Spanish and 6 for Italian. Because retrieval effectiveness depends upon translation, queries of 2-3 words cause poor retrieval performance even if translation of just one word is wrong. For example, topic 25 performs better across all languages with document translation. This is because the word for “canal” is mis-translated in all languages to “channel” in English, but correctly

[6] McCarley [6] also found this to be true for French–English.

Table 2: A summary of retrieval effectiveness for query translation

	MAP	%English MAP	Recall	Avg $P_{norm100}$	Topics good	Topics bad
de_en	0.5047	81.6%	0.8408	0.3541	6	1
de_en_fb	0.4994	80.7%	0.8251	0.3409	7	2
fr_en	0.4516	73.0%	0.7768	0.3422	6	3
fr_en_fb	0.4567	73.8%	0.8613	0.3297	6	4
it_en	0.4934	79.8%	0.7648	0.3668	4	1
it_en_fb	0.4507	72.9%	0.7153	0.3167	7	2
es_en	0.4654	75.2%	0.8842	0.3484	5	1
es_en_fb	0.4718	76.3%	0.7503	0.3408	6	3

Table 3: A comparison of the Sheffield results with other submissions

	Rank Position	
Language	With Submitted	With Highest Result
English	3	1
Italian	2	1
German	13	4
French	4	4
Spanish	1	1

translated from English into the four target languages (e.g. "canal"→"kanale"→"channel"). This shows that translation for this word is not symmetric, i.e. that English→X $\neq$ X→English.

Of course, in some cases query translation is better than document translation. For example, German topic 20 ("river with a viaduct in the background") performs badly for document translation because crucial words are not translated, e.g. "viaduct", or words are mis-translated altogether. Query translation fails because either words are mis-translated (e.g. "boats in a channel" rather than "boats in a canal"), not translated at all (e.g. "External views of Egyptian templi" rather than "Exterior views of Egyptian temples") , or synonymous terms are used instead which are correct but do not match the caption terms (e.g. "images of English beacons" rather than "images of English lighthouses").

We find the following correlations between average precision scores for query and document translation (using Spearman's rho with $p < 0.01$**): German (0.123), Spanish (0.511**), French (0.736**) and Italian (0.725**). The last three languages show a significant correlation between average precision using either document or query translation (i.e. topics perform similarly using either approach). German is not correlated because of reasons given previously for query and document translation.

4 Comparison with other ImageCLEF submissions

Table 3 summarises the results obtained against all others submitted to ImageCLEF 2004. I submitted results for document translation with relevance feedback only assuming these would be the highest results, but this did not prove to be true. The average rank position across all languages using the submitted results is 4.6; whereas using the highest results an average position of 2.2. is obtained. It is somewhat surprising that the Spanish submission using document translation proved to be the highest because our approach used very little language processing and knowledge of translation. With query translation and no feedback, we could have improved the rank position of our English, Italian and German entries.

5 Conclusions and Future Work

In this paper I have presented experiments whereby I compare document and query translation using the `Systran Professional` and `Babelfish` MT systems for the ImageCLEF 2004 ad hoc image retrieval task. On average query translation outperforms document translation for Spanish, Italian, French and German texts, but this varies across both language and topic. Various translation errors cause low retrieval effectiveness for both document and query translation methods. Given the effort involved in document translation and lower retrieval performance than query translation, it would appear that the latter approach is better for this retrieval task. Document translation can be applied after retrieval prior to presenting captions to the user rather than introducing errors into the retrieval process.

However, we observe some interesting effects across individual topics where document translation outperforms query translation. This is particularly true when queries are short and crucial query terms are mis-translated or not translated at all. Because caption translation is feasible for image collections because captions are typically much shorter than most other texts, further work could explore combining both document and query translation to improve results and take advantage of which method gives better retrieval performance by merging together results from document and query translation. I would also like to explore improving the feedback model by training parameters for optimal values. Finally, I would like to explore methods to improve document translation into German by making the texts cleaner prior to running the MT system.

6 Acknowledgements

This work was carried out within the Eurovision project at Sheffield University funded by the EPSRC (Eurovision: GR/R56778/01).

References

[1] Grefenstette, G.: Cross Language Information Retrieval. Kluwer Academic Publishers, Norwell, MA, USA (1998)

[2] Peters, C., Braschler, M.: Cross language system evaluation: The clef campaigns. Journal of the American Society for Information Science and Technology (2001) 1067–1072

[3] Oard, D.: Serving users in many languages. D-Lib magazine (1997)

[4] Clough, P., Sanderson, M.: User experiments with the eurovision cross-langauge image retrieval system. In: JASIST, in submission. (2004)

[5] Oard, D.: A comparative study of query and document translation for cross-language information retrieval. In: Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas. (1998) 472–483

[6] McCarley, S.: Should we translate the documents or the queries in cross language information retrieval? In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. (1999) 208–214

[7] Fujii, A.and Ishikawa, T.: Applying machine translation to two-stage cross-language information retrieval. In: Proceedings of the 4th Conference of the Association for Machine Translation in the Americas (AMTA-2000). (2000) 13–24

[8] Clough, P., Sanderson, M., Müller, H.: The cross language image retrieval track (imageclef) 2004. In: Submission, to appear. (2004)

[9] Zhai, C., Lafferty, J.: A study of smoothing methods for langauge models applied to ad hoc information retrieval. In: Proceedings of SIGIR'2001. (2001) 334–342

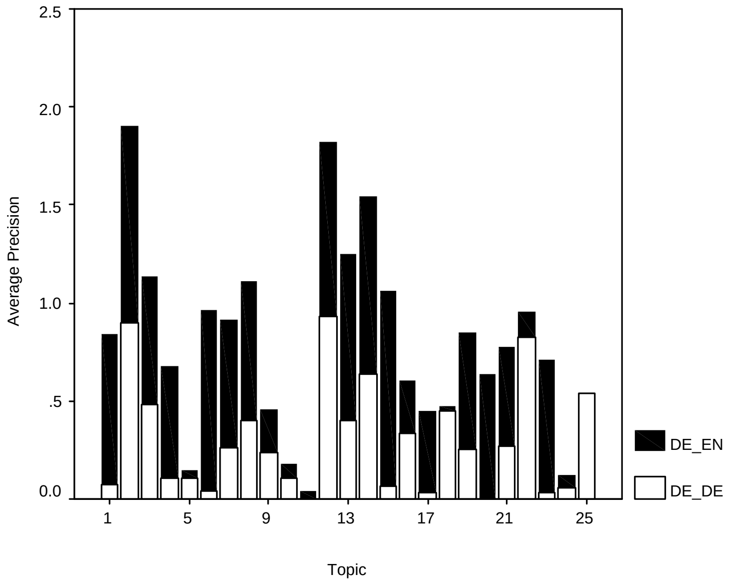

Figure 3: Average Precision for German topics

[10] Clough, P., Sanderson, M.: Assessing translation quality for cross language image retrieval. In: Submission, to appear. (2003)

[11] Hutchins, W., Somers, H.: An Introduction to machine Translation. Academic Press, London, England (1986)

[12] Clough, P., Sanderson, M.: The effects of relevance feedback in cross language image retrieval. In: Proceedings of the 26th European Conference on Information Retrieval (ECIR'04). (2004) 353–363

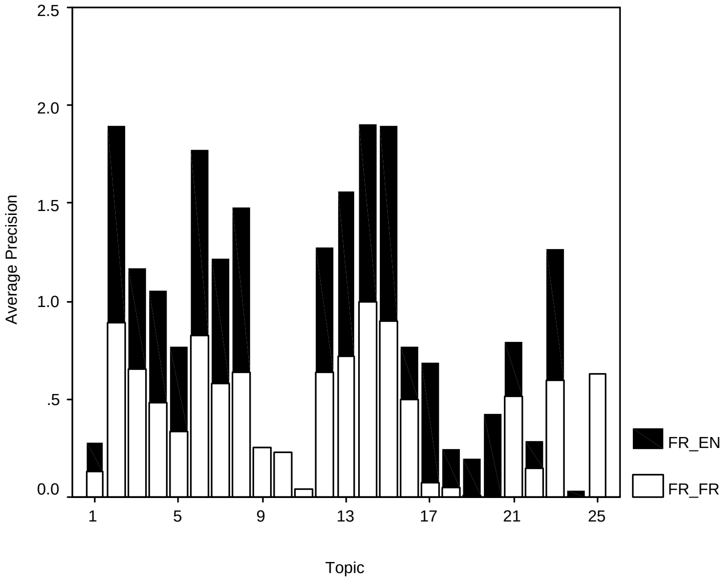

Figure 4: Average Precision for French topics

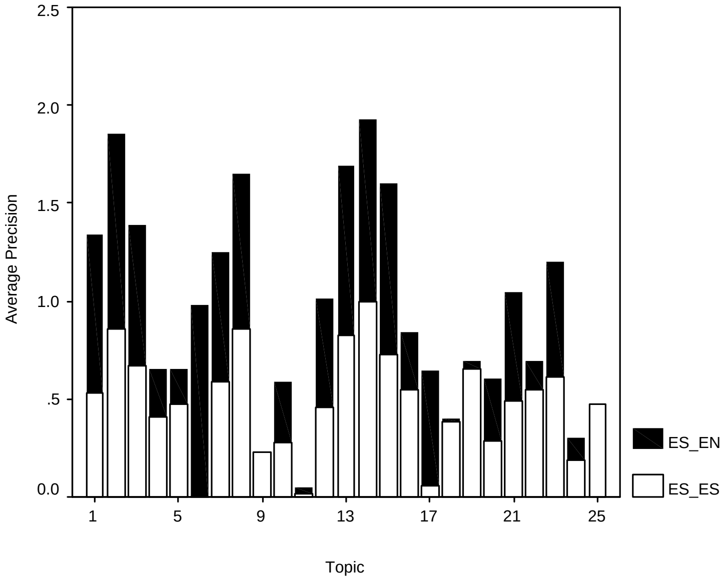

Figure 5: Average Precision for Spanish topics

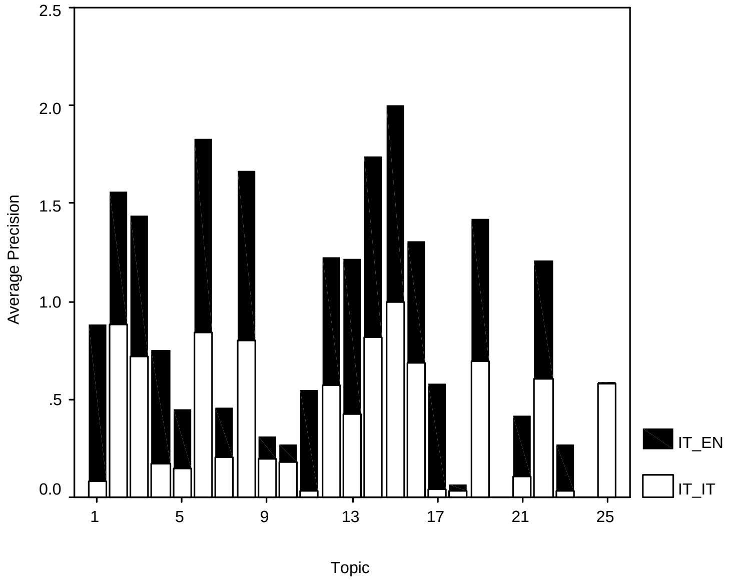

Figure 6: Average Precision for Italian topics