UNED at iCLEF 2003: Searching Cross-Language Summaries

Fernando López-Ostenero, Julio Gonzalo, Felisa Verdejo
Departamento de Lenguajes y Sistemas Informáticos, UNED
{*flopez,julio,felisa*}*@lsi.uned.es*

Abstract

The UNED phrase-based cross-language summaries were first introduced at iCLEF 2001 as a translation strategy which permitted faster document selection with roughly the same accuracy than full Machine Translation. For our iCLEF 2003 participation, we test the validity of our summaries as cross-language indexes for the retrieval stage of the interactive search process. We compare a reference system that performs query translation (and then retrieves target-language documents) with a proposed system that directly retrieves cross-language summaries with the source-language query. The performance of both systems is very similar, confirming that UNED summaries are viable for cross-language indexing. This approach is trivially scalable to more than one target language, opening an interesting path for truly multilingual search assistance.

1 Introduction

For Cross-Language Information Retrieval (CLIR) purposes, query translation is normally preferred to document translation due to the high computational cost of performing Machine Translation (MT) on a whole document collection. Still document translation as, at least, two clear advantages over query tranlation: first, translation can be done more accurately, because there is more context; and second, that merging of ranked results from different target languages is not necessary, because there is only one retrieval rank in the user's language. In any case, the increase in computational cost (comparing to query translation) is too dramatic to turn document translation into a mainstream approach.

From an interactive point of view, however, document translation becomes an attractive option:

- When the user of a CLIR system comes into play, some form of document translation becomes unavoidable: the user needs some indication about the content of foreign-language documents in order to determine a) the relevance of the documents retrieved by the system, b) whether there is a need to refine the query and c) if so, which terms should be added/removed from the original query.
- If the user is involved in the query translation process, the interaction can become too complex when there is more than one target language involved. Moreover, previous research has shown that the users prefer not to be involved in the query translation process [5, 1]. With a document translation approach to interactive CLIR, both problems simply dissappear.

Therefore, in interactive CLIR the question is no longer whether we should translate documents or not, but *how* to translate documents in a way that *a*) facilitates searching tasks to the user (document selection and query refinement); *b*) minimizes the computational cost of translation, and *c*) in the case of a document translation approach, that provides optimal material for the automatic retrieval stage in the user's language.

In previous iCLEF editions, we have proposed a method of translation that reasonably satisfies requisites *a* and *b*. The method generates summarized translations of target-language documents

which rely essentially on noun phrase extraction and translation. While rather crude from a Computational Linguistics point of view, such summaries have excellent features for Multilingual Information Access:

- In iCLEF 2001, we obtained quantitative evidence that such cross-language summaries could be better for document translation purposes than full MT (users judgements were equally precise with both approaches, but summaries permitted faster judgments).
- In iCLEF 2002, our noun-phrase based summaries proved to be useful as a basis for query formulation and refinement.
- Phrase-based summaries contain only 30% as many words as the original documents, and can be generated more than one order of magnitude faster than full machine translations.

It seems, then, reasonable to think of phrase-based summaries as good candidates for a document translation approach to Multilingual Information Access. This is the hypothesis that we seek to test in our iCLEF 2003 experiment. The challenge is performing retrieval with phrase-based summaries as the only source of query-language indexes for target-language documents, because the size of the index set will be just one third of its monolingual or MT counterparts.

In Section 2, we review the main features of phrase-based summaries. In Section 3 we discuss the experiment design; in Section 4 we present the results of our experiment, and finally we draw some conclusions.

2 UNED Cross-language summaries

UNED approach to *Cross-Language Search assistance* (defined as the problem of assisting a user to search and detect relevant documents in a foreign-language text collection) is based on noun phrases as fundamental units for translation (either query or document translation) and formulation of user needs.

Cross-language pseudo summaries are an essential part of the approach. They simply consist in the list of noun phrases present in the document, listed in order of appearance, and translated according to a simple greedy algorithm that makes use of a database of bilingual alignments between two and three-lemma phrases in the source and target languages [2]. The phrase alignment resource is built using a simple noun phrase extractor [4] on two comparable text collections (EFE 94 and LA Times 94 in our case) and an alignment algorithm based on co-occurrence of candidate translations (via bilingual dictionaries) and phrase frequency measures. The algorithm produces sets of phrases which are assumed to be *equivalent under translation*, and the most frequent phrase in each set is said to be the *canonical translation* for each member of the equivalent set (see Figure 1 for an example).

SPANISH
acuerdo de libre comercio
acuerdos de libre comercio
acuerdo libre comercio
acuerdo de libre cambio
acuerdos de libre cambio
convenio de libre comercio
convenios de libre comercio
compromiso de libre comercio
...

ENGLISH
free trade agreement
free trade accord
free trade pact
free trade beyond the pact
free trade pacts
free trade agreements
free trade arrangements

Figure 1: Example of sets of noun phrases equivalent under translation.

Table 2 lists the size of the bilingual noun phrase alignments as extracted from EFE 1994 and LA Times 1994 corpora. In its current version, the alignment algorithms takes 17 hours to generate two-lemma alignments (8 hours of preprocessing and 9 hours of alignment) plus 60 hours to generate three-lemma alignments.

Translation of non-aligned noun phrases (including phrases with more than three lemmas) is done with a greedy algorithm that translates, at each step, the two or three-lemma sub-phrase which has a better translation under the alignment resource, and uses overlapping phrases to translate the remaining words taking the context into account. The algorithm is described in detail in [2]. The translation of the whole LA Times 1994 collection (approx. 110.000 documents) takes a total of 14 hours (12 hours for summarization and 2 hours for translation). The average size of a summary is 30% the size of the original document.

A comparative example of translations provided by Systran and by our system can be seen in the Appendix.

Language	**2 lemmas**	**3 lemmas**
English	1,700,183	288,872
Spanish	1,953,849	347,920

Figure 2: Aligned phrases resource

3 Experiment design

The goal of the experiment is testing whether searching cross-language summaries with the original query can match searching the original documents with a translated version of the query, in an interactive CLIR setting. As reference system, we have chosen the best of the two approaches tested in our iCLEF 2002 experiment. In this approach, users interact with the system to formulate an optimal query as a set of noun phrases. Then, query translation is performed automatically (via the database of aligned phrases) and the retrieved set of documents can be examined via cross-language summaries. As the user do not have to deal with foreign-language expressions at any time, the translation and retrieval steps can be substituted for a direct retrieval on document translations without altering the interface with the user. This is very convenient for our experiment, because it permits a direct comparison of query translation versus document translation strategies without any additional interference.

A detailed description of both systems to be compared follows:

1. **Initial query formulation**: the user reads the topic description and formulates an initial query freely. The time for reading and typing this initial query is not computed as searching time.

2. **Query formulation by phrases**: the system suggests a maximum of 10 phrases related to the initial user query. The user can either a) select a number of them and perform the initial search, or b) type in some additional words and ask the system to recompute the phrase suggestions.

3. **Document retrieval**:

   - In the reference system (**Query Translation**) , phrases are translated into English via the phrase alignment dictionary, and then a search is performed against the LA Times collection.
   - In the contrastive system (**Document translation**), the original Spanish phrases are used to search the collection of Spanish summaries of LA Times documents.

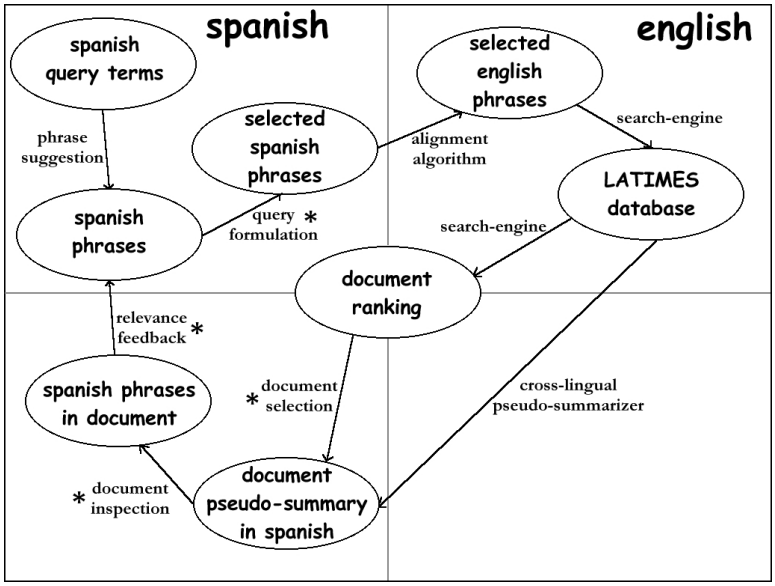
4. **Document ranking**: The result of the search is a ranked list of LA Times documents, with a colour code to indicate whether each document has been judged as relevant, not relevant, unsure, or has not been judged yet. Each document is displayed in the ranked list as a Systran translation of its title.

5. **Document selection**: When the user clicks on a document title, its Spanish summary is shown to the user. The document can then be judged as relevant, not relevant or unsure.

6. **Query refinement**: There are two ways of refining the query:

   - *Phrase feedback*: If the user clicks on a phrase inside a document summary, the phrase is added to the query and the document ranking is updated with the enhanced query.
   - *Direct reformulation*: at any point during the search, the user can select/deselect additional phrases to the query, and can introduce new words to the phrase suggestion window.

Although the difference between both systems is transparent to the user, the architecture and implications of each of them are quite different. Figures 3 and 4 compare both approaches visually. The document translation approach can be trivially extended to more than one document language, as can be seen in Figure 5. Using phrase-based summaries, the document translation approach can be applied to a multi-language collection increasing its size in only 30% with respect to the original size per user language considered. For instance, a collection with four document/user languages would only double its size under this document translation approach. This is a challenge for the retrieval phase, because the set of indexes is much smaller than the original. Our hypothesis is that any possible difference in the quality of the rankings will not have an appreciable impact on the interactive searching task.



("*" indicates interaction with the user)

Figure 3: Reference system: query translation approach

We have used eight native Spanish searchers for our experiment, the LA Times 1994 collection as the document set, and the eight official iCLEF topics (in Spanish) extracted from the CLEF 2002 set. This year, our searchers do not interact with English at any moment, hence we were not specially cautious with English proficiency when recruiting volunteers. We could then focus on recruiting searchers with long experience using search engines (something that previous years was not granted). In the end, all eight users have medium English skills (which are not put into practice in the experiment) and are highly experienced in web searching.

As specified in the track guidelines, each search session consisted on a unique user/system/topic combination according to a latin-square matrix design [3] that sorts out the differences between individual topics (some topics are easier than others), individual users (some users are better than others) so that a difference between systems can be established without bias effects. Both systems are identical to the user, therefore only one previous training phase was needed. Each user then performed eight searches (on the eight iCLEF topics), with a limit of 10 minutes per search. Their

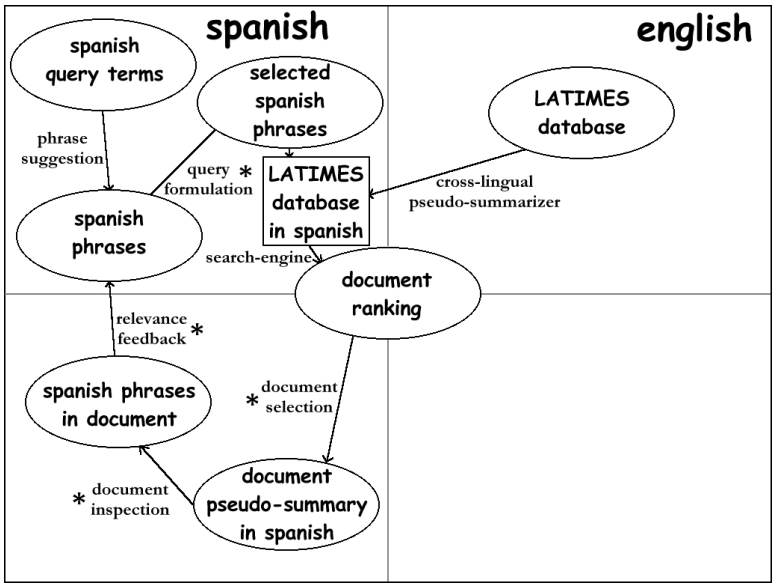


Figure 4: Contrastive system: document translation approach

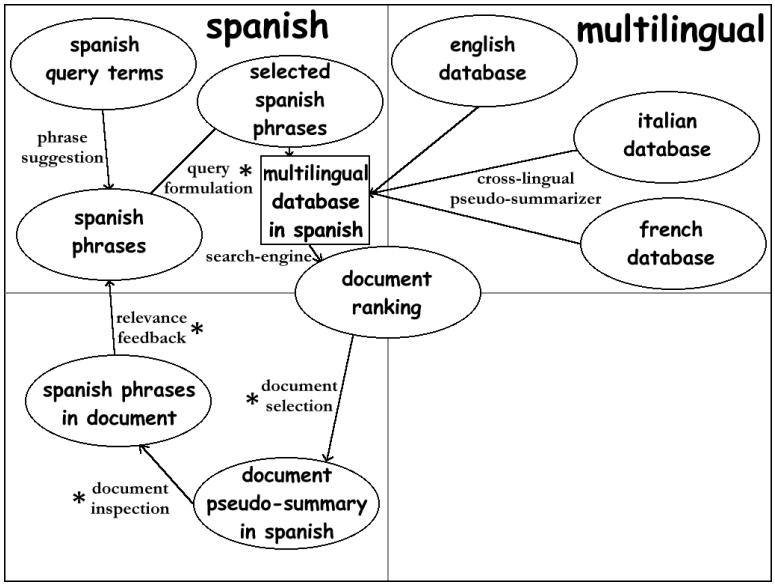


Figure 5: Multilingual document translation approach

goal was to retrieve as many relevant documents as possible, focusing on precision rather than recall (it is more important that the selected documents are actually relevant, than finding every relevant item for a given topic).

4 Results and discussion

4.1 Official results

The official Van Rijsbergen's $F_\alpha = 0.8$ measure ($\alpha = 0.8$ favors precision rather than recall) can be seen in Figure 6. Both systems receive the same score, confirming our hypothesis that phrase-based summaries can be used for a cross-language indexing of the collection. Precision and recall are also almost identical for both approaches.

System	Precision	Recall	F_α
Query translation	.51	.14	.29
Document translation	.53	.14	.29

Figure 6: Official UNED results

4.2 Dependence on user and topic

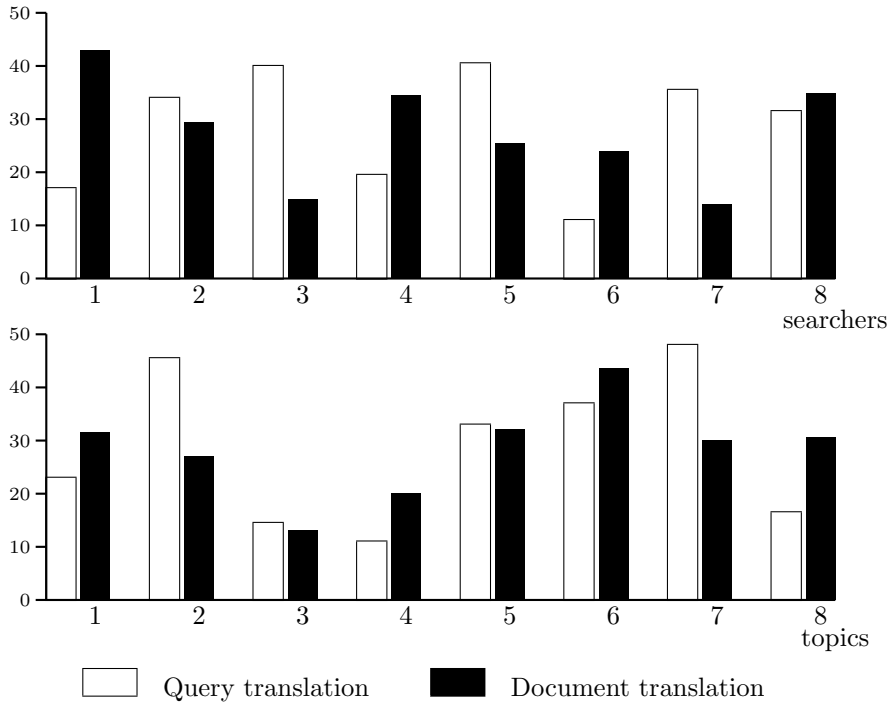


Figure 7: F_α results by searcher and topic

Figure 7 separates results by searcher and by topic. Topics vary in difficulty as expected, being topics 3 and 4 harder than the others. There is also variability between users, but part of this variability can be explained by the distribution of topics to systems assigned to each searcher. For instance, users 1 and 3 have opposite results (query translation is much worse for user 1 and much better for user 3), but they have an inverse assignment of topics and systems, and the two difficult

topics are done with different systems from user 1 to user 3. Overall, the detailed results are a good sample of how the latin square design filters out possible topic/system/user combination bias.

4.3 Initial Precision

Figure 8 shows the initial precision, measured on the twenty first documents retrieved by the system after initial query formulation by the user. Remarkably, the query translation approach performs 40% better, indicating that the quality of batch retrieval might be better with a full document index than with an index based on the cross-language summary. But this initial difference is not reflected in the final retrieval results, suggesting that other interactive factors are predominant in the overall search results (for instance, facilities for query refinement).

System	Initial precision
Document translation	10%
Query translation	14% (+40%)

Figure 8: Average initial precision

4.4 Selections across time

Figure 9 shows the overall number of documents selected across the 10 minute searching time. Unlike our previous iCLEF 2002 experiment, this year there are no sensible differences between systems. In spite of the initial advantage in precision, the pattern of selections across time is basically equal for both systems. The growing curve of selections suggest that results would have improved with longer searches.

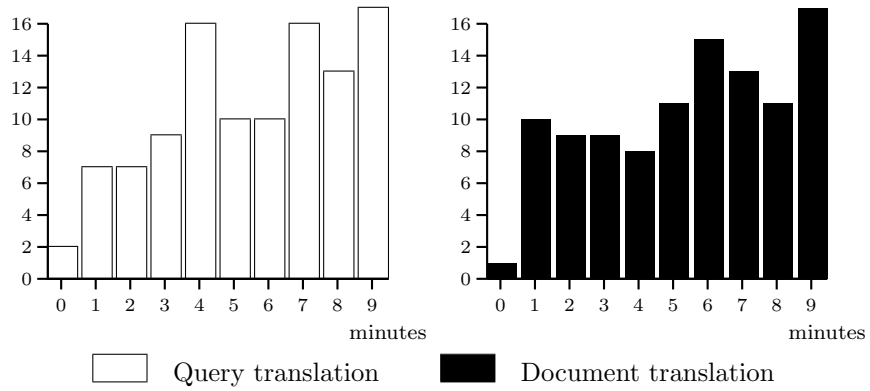


Figure 9: Number of selected documents across time

5 Conclusions

Document translation can be a viable approach to Multilingual Information Access once we find simplified, task-oriented ways of translating documents without the computational cost associated to commercial Machine Translation systems. In this experiment, we have proved that phrase-based summaries, although rather crude from a pure MT point of view, can be successfully used for Cross-Language searching.

Phrase-based summaries can be generated much faster than full machine translations, and occupy only 30% of disk space. In these conditions, phrase-based summaries can be used to benefit from the main advantages of document-translation approaches to interactive CLIR:

- At document selection time, translations do not have to be generated on the fly, because all documents have been previously translated for indexing. Hence the interactive search process is not retarded by on-line translation.
- In a truly multilingual setting (with more than one target language), the complexity of translating the query to several languages (a big impediment if query translation is done interactively) and the problem of merging ranked results from different languages dissappear.

In our present experiment, query formulation is somewhat restrained by the fact that the user has to formulate his/her query as a set of noun phrases that can in turn receive an appropriate automatic translation via the alignment resource. This makes perfect sense in the reference (query translation) system, as proved in our iCLEF 2002 experiment. But it might be an excessive constraint in the document selection approach. We plan to experiment whether a more classical, monolingual search interface (with the possibility of adding free terms at any moment of the search process) might even improve the results obtained by query translation approaches to foreign-language search assistance.

Appendix: Translation example

This is an example of the translation of a short LA Times document, both with Systran Professional 3.0 (as provided to iCLEF participants) and with phrase-based summaries.

Original document

WORLD CUP SOCCER '94 / THE FIRST ROUND; SPOTLIGHT; NOT AGAINST BRAZIL

Reuters news service sent a picture of Carlos Alberto Torres, the captain of Brazil's 1970 World Cup championship team, talking with Lothar Matthaeus, captain of Germany's team, at a recent practice at Southern Methodist University. The caption information included with the photo identified Torres as a German fan. ELLIOTT ALMOND

Systran translation

FÚTBOL '94 / EL PRIMER REDONDO DE LA TAZA DEL MUNDO; PROYECTOR; NO CONTRA EL BRASIL

El servicio de noticias de Reuters envió un cuadro de Carlos Alberto Torres, el capitán de Equipo 1970 del campeonato de la taza del mundo del Brasil, hablando con Lothar Matthaeus, capitán del equipo de Alemania, en una práctica reciente en Methodist meridional Universidad. La información del subtítulo incluida con la foto identificó a Torres como ventilador alemán. ALMENDRA DE ELLIOTT

Cross-lingual pseudo-summary

copa del mundo de fútbol
primera ronda

reuters **news service**
ver carlos
campeón de la copa del mundo equipo
lothar matthaeus
práctica del reciente universidad metodista del sur
pie informacion
torres **aficionados alemanes**
elliott almendra

Acknowledgments

We are indebted to Anselmo Peñas for the provision of the WTB phrase extraction software. This research has been funded by the Spanish Government, project *Hermes* (TIC2000-0335-C03-01).

References

[1] F. López-Ostenero, J. Gonzalo, A. Peñas, and F. Verdejo. Phrases are better than words for interactive cross-language query formulation and refinement. In *Evaluation of Cross-Language Information Retrieval Systems, Springer-Verlag Lecture Notes in Computer Science*, 2002.

[2] Fernando López-Ostenero. *Un sistema interactivo para la búsqueda de información en idiomas desconocidos por el usuario*. PhD thesis, UNED, 2002.

[3] Douglas W. Oard and J. Gonzalo. The clef 2003 interactive track. In *Proceedings of CLEF 2003*, 2003.

[4] Anselmo Peñas, Julio Gonzalo, and Felisa Verdejo. Cross-language information access through phrase browsing. In *Applications of Natural Language to Information Systems*, Lecture Notes in Informatics, pages 121–130, 2001.

[5] D. Petrelli, M. Beaulieu, M. Sanderson, G. Demetriou, and P. Herring. Is query translation a distinct task from search? In *Proceedings of CLEF 2002*, 2002.

F_α measure

Searcher \ Topic	1	2	3	4	5	6	7	8	Avg.
1	0.13	0.32	0	0.24	**0.43**	**0.48**	**0.36**	**0.45**	0.3
2	**0.44**	**0.43**	**0.21**	**0.11**	0.37	0.43	0.56	0	0.32
3	**0.45**	**0**	**0.16**	**0**	0.56	0.37	0.68	0	0.28
4	0.09	0.33	0.16	0.2	**0**	**0.37**	**0.56**	**0.45**	0.27
5	0.58	**0.65**	**0.16**	0	**0.22**	0.37	0.68	**0**	0.33
6	0.13	**0**	**0**	0	**0.64**	0.32	0	**0.33**	0.18
7	**0.11**	0.53	0	**0.3**	0.22	**0.43**	**0**	0.67	0.28
8	**0.26**	0.65	0.43	**0.4**	0.18	**0.46**	**0.28**	0	0.33
Avg.	0.27	0.36	0.14	0.16	0.33	0.4	0.39	0.24	0.29

Precision

Searcher \ Topic	1	2	3	4	5	6	7	8	Avg.
1	0.17	0.3	0	0.4	**1**	**1**	**1**	**0.5**	0.55
2	**0.6**	**0.43**	**0.4**	**0.17**	0.67	0.75	1	0	0.5
3	**1**	**0**	**1**	**0**	1	1	1	0	0.63
4	0.1	0.5	1	1	**0**	**1**	**1**	**0.5**	0.64
5	1	**0.75**	**1**	0	**0.5**	1	1	**0**	0.66
6	0.17	**0**	**0**	0	**1**	0.67	0	**0.33**	0.27
7	**0.13**	0.67	0	**0.67**	0.5	**0.75**	**0**	0.67	0.42
8	**0.27**	0.75	0.8	**0.75**	0.33	**0.67**	**0.5**	0	0.51
Avg.	0.43	0.42	0.53	0.37	0.62	0.85	0.69	0.25	0.52

Recall

Searcher \ Topic	1	2	3	4	5	6	7	8	Avg.
1	0.07	0.43	0	0.1	**0.13**	**0.16**	**0.1**	**0.33**	0.16
2	**0.21**	**0.43**	**0.07**	**0.05**	0.13	0.16	0.2	0	0.16
3	**0.14**	**0**	**0.04**	**0**	0.2	0.11	0.3	0	0.1
4	0.07	0.14	0.04	0.05	**0**	**0.11**	**0.2**	**0.33**	0.12
5	0.21	**0.43**	**0.04**	0	**0.07**	0.11	0.3	**0**	0.14
6	0.07	**0**	**0**	0	**0.27**	0.11	0	**0.33**	0.1
7	**0.07**	0.28	0	**0.1**	0.07	**0.16**	**0**	0.67	0.17
8	**0.21**	0.43	0.15	**0.14**	0.07	**0.21**	**0.1**	0	0.16
Avg.	0.13	0.27	0.04	0.05	0.12	0.14	0.15	0.21	0.14

Figure 10: Detailed results by topic and searcher