Noun Phrase Translations for Cross-Language Document Selection

Fernando López-Ostenero, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo

Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
E.T.S.I Industriales, Ciudad Universitaria s/n, 28040 Madrid, SPAIN
{flopez,julio,anselmo,felisa}@lsi.uned.es
http://sensei.lsi.uned.es/NLP

Abstract. This paper presents results for the CLEF Interactive Cross-Language Document Selection task at the UNED. Two translation techniques were compared: the standard Systran translations provided by the CLEF organizers as a baseline, and a phrase-based pseudo-translation approach that uses a phrase alignment algorithm based on comparable corpora. The hypothesis being tested was that noun phrase translations could serve as summarized information for relevance judgment without compromising the precision of such judgments. In addition, we wanted to have an indirect measure of the quality of our phrase extraction process, that had been previously developed for an interactive CLIR application. The results of the experiment confirm that the hypothesis is reasonable: a set of 8 monolingual Spanish speakers judged English documents with the same precision for both systems, but achieved 52% more recall using phrasal translations than using full Systran translations.

1 Introduction

The goal of the CLEF 2001 interactive track (iCLEF) was to compare ways of informing a monolingual searcher about the content of documents written in foreign languages: a better system will allow for better relevance judgments and therefore better foreign-language document selection [2]. The baseline approach is using standard Machine Translation (MT) to produce translated versions of the documents.

Our intuition was that translations produced by MT are noisy and much harder to read and understand than hand-written documents. Perhaps a smaller amount of information, with the best translated phrases highlighted, could facilitate relevance judgment without a significant loss of precision.

To test such a hypothesis, we took advantage of a phrase extraction software previously developed within our research group for an interactive CLIR application [3]. This software is able to index noun phrases in large text collections in a variety of languages (including Spanish and English), providing a good starting material for a phrase-based summarized translation of the documents used in the iCLEF task. Then we performed the following steps:

C.A. Peters et al. (Eds.): CLEF 2001, LNCS 2406, pp. 320–331, 2002.

1. Extract phrasal information from the 200 documents (50 per iCLEF query) of the English CLEF 2000 collection.
2. Find a (large) Spanish corpus comparable with the iCLEF documents. This choice was easy, as the CLEF 2001 test set includes a comparable collection (EFE newswire 1994) of 250,000 Spanish documents (approximately 1Gb of text including SGML tags).
3. Extract phrasal information from the EFE 1994 collection.
4. Develop an alignment algorithm to obtain optimal Spanish translations for all phrases in the English documents.
5. Incorporate phrasal translations in a display strategy for the iCLEF document selection task.
6. Carry out the comparative evaluation between our system and Systran translations, following the iCLEF 2001 guidelines.

Besides testing our main hypothesis, we had three additional goals: first, scaling up the phrase extraction software to handle CLEF-size collections; second, enriching such software with a phrase-alignment algorithm that exploits comparable corpora; and third, obtaining an indirect measure (via document selection) of the quality of that software.

In Section 2, we describe our phrase-based approach to document translation. In Section 3, the experimental setup for the evaluation is explained. In Section 4, results are presented and discussed. Finally, in Section 5 we draw some conclusions.

2 Phrase-Based Pseudo-Translations

2.1 Phrase Extraction

We have used the phrase extraction software from the *UNED WTB Multilingual search engine* [3]. This software performs robust and efficient noun phrase extraction in several languages, and provides two kinds of indexes:

- an index that maps every (lemmatized) word to every noun phrase that contains a morphological variant of the word, and
- an index that maps every noun phrase to documents that contain that phrase.

Noun phrases are extracted using shallow NLP techniques:

1. Words are lemmatized using morphological analyzers. The Spanish processor uses MACO+ [1], and the English processor uses TreeTager [4].
2. Words are tagged for Part-Of-Speech (POS). No POS tagger, to our knowledge, is able to process gigabytes of text. Therefore, a fast approximation to tagging is performed: in the case of Spanish, a set of heuristics has been devised to ensure maximal recall in the phrase detection phase. For other languages, the most frequent POS is assigned to all occurrences of a word.

3. A shallow parsing process identifies noun phrases that satisfy the following (flexible) pattern:

$$[noun|adj][noun|adj|prep|det|conj]^*[noun|adj]$$

4. Finally, indexes for *lemma→phrases* and *phrase→documents* are created.

The collection of 200 English documents is very small and poses no problem for indexing. The EFE collection, however, consists of about 250,000 documents corresponding to about 1Gb of text. Before attempting this iCLEF experiment, the largest collection processed with our system contained 60,000 documents. In order to process the EFE collection with our (limited) hardware resources, it was necessary to re-program most of the system.

These are the approximate figures for the indexing process: 375,000 different words were detected, from which 250,000 were not recognized by the morphological analyzer, because they correspond to proper nouns, typos, foreign words, or words not covered by the dictionary.

Overall, 280,000 different lemmas (including unknown words) are considered, and 26,700,000 different candidate phrases are detected. From this set, we have retained the 3,600,000 phrases that appear more than once in the collection.

In the WTB search engine, such indexes are used to provide multilingual phrase-browsing capabilities in an interactive CLIR setting. In the present work, however, this data is used as statistical information to provide translations for English phrases in iCLEF documents.

2.2 Phrase Alignment

For each English phrase, we start translating all content words in the phrase using a bilingual dictionary. For instance:

```
phrase:       "abortion issue"
lemmas:       abortion, issue
translations: abortion -> aborto
              issue    -> asunto, tema, edición, número, emisión,
                          expedición, descendencia, publicar,
                          emitir, expedir, dar, promulgar
```

For each word in the translations set, we consider all Spanish phrases that contain that word. The set of all phrases forms the *pool of related Spanish phrases*.

Then we search all phrases that contain only (and exactly) one translation for every term of the original phrase. This subset of the Spanish related phrases forms the *set of candidate translations*. In the previous example, the system finds:

		phrase	frequency
		tema del aborto	16
		asunto del aborto	12
abortion issue	⇒	asuntos como el aborto	5
		asuntos del aborto	2
		temas como el aborto	2
		asunto aborto	2

If the subset is non-empty (as in the example above), the system selects the most frequent phrase as the best phrasal translation. Therefore *"tema del aborto"* is (correctly) chosen as translation for *"abortion issue"*. Note that all other candidate phrases also disambiguated *"issue"* correctly as *"tema, asunto"*.

Other alignment examples include:

English	# candidates	selected	frequency
abortion issue	6	tema del aborto	16
birth control	3	control de los nacimientos	8
religious and cultural	10	culturales y religiosos	14
last year	52	año pasado	8837

The most appropriate translation for *"birth control"* would rather be *"control de la natalidad"* (with a frequency of 107), but the dictionary does not provide a link between *"birth"* and *"natalidad"*. The selected term *"control de los nacimientos"*, however, is unusual but understandable (in context) for a Spanish speaker.

If the set of candidate translations is empty, two steps are taken:

1. **Subphrase translation**: the system looks for maximal sub-phrases that can be aligned according to the previous step. These are used as partial translations.
2. **Word by word contextual translation**: The remaining words are translated using phrase statistics to take context into account: from all translation candidates for a word, we choose the candidate that is included in more phrases from the original pool of related Spanish phrases.

For instance:

```
phrase:  "day international conference on population and development"
lemmas:    day, international, conference, population, development
possible translations:
    day           -> día, jornada, época, tiempo
    international -> internacional
    conference    -> congreso, reunión
    population    -> población, habitantes
    development   -> desarrollo, avance, cambio, novedad, explotación,
                     urbanización, revelado
```

```
subphrase alignments:
    day international            -> jornadas internacionales
    day international conference -> jornada del congreso internacional
word by word translations:
    population                   -> población
    development                  -> desarrollo
final translation:
   "jornada del congreso internacional población desarrollo"
```

Note that, while the indexed phrase is not an optimal noun phrase (*"day"* should be removed) and the translation is not fully grammatical, the lexical selection is accurate, and the result is easily understandable for most purposes (including document selection).

2.3 Phrase-Based Document Translation

The pseudo-translation of the document is made using the information obtained in the alignment process. The basic process is:

1. Find all maximal (i.e., not included in bigger units) phrases in the document, and sort them by order of appearance in the document.
2. List the translations obtained for each original phrase according to the alignment phase, highligting:
   - Phrases that have an optimal alignment (boldface).
   - Phrases containing query terms (bright colour).

As an example, let us consider this sentence from one of iCLEF documents:

> English sentence
>
> the abortion issue dominated the nine-day International Conference on Population and Development.

A valid manual translation of the above sentence would be:

> Manual translation
>
> el tema del aborto dominó las nueve jornadas del Congreso Internacional sobre Población y Desarrollo.

while Systran produces:

> Systran MT translation
>
> la edición del aborto dominó el de nueve días Conferencia internacional sobre la población y el desarrollo.

Aside from grammatical correctness, Systran translation only makes one relevant mistake, interpreting *"issue"* as in *"journal issue"* and producing *"edición del aborto"*(meaningless) instead of *"tema del aborto"*.

Our phrase indexing process, on the other hand, identifies two maximal phrases:

abortion issue
day International Conference on Population and Development

which receive the translations showed in the previous section. The final display of our system is:

> PHRASAL PSEUDO-TRANSLATION
>
> **tema del aborto**
> **jornada del congreso internacional** población desarrollo

where boldface is used for optimal phrase alignments, which are supposed to be less noisy translations. If any of the phrases contain a (morphological variant of a) query term for a particular search, the phrase is further highlighted.

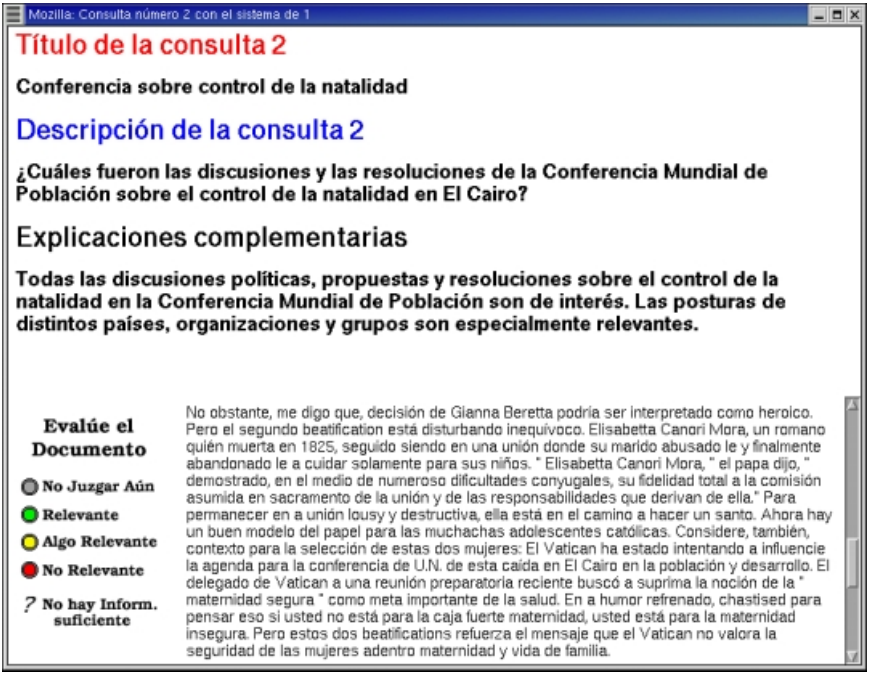

Fig. 1. Search interface: MT system

3 Experimental Setup

3.1 Experiments and Searchers

We conducted three experiments with different searcher profiles: for the main experiment, we recruited 8 volunteers with low or no proficiency at all in the English language. For purposes of comparison, we formed two additional 8-people groups with mid-level and high-level English skills.

3.2 Search Protocol and Interface Description

We followed closely the search protocol established in the iCLEF guidelines [2]. The time for each search, and the combination of topics and systems, were fully controlled by the system interface. Most of the searchers used the system locally, but five of them (UNED students) carried out the experiments via Internet from their study center (in the presence of the same supervisor).

Figure 1 shows an example of a document displayed using the Systran MT system. Figure 2 shows the same document paragraph using our phrase-based system. The latter shows less information (only noun phrases extracted and translated by the system), highlights phrases containing query terms (bright green) and emphasizes reliable phrasal translations (boldface).

4 Results and Discussion

The main precision/recall and $F\alpha$ figures can be seen in Table 4. In summary, the main results are:

- In the main experiment with monolingual searchers ("Low level of English"), precision is very similar, but phrasal translations obtain 52% more recall. Users judge documents faster without loss of accuracy.
- Users with good knowledge of English show a similar pattern, but the gain in recall is lower, and the absolute figures are higher both for MT and phrasal translations. As unknown words remain untranslated and English-speaking users may recognize them, these results are coherent with the main experiment. See Figure 3 for a comparison between low and high English skills.
- Mid-level English speakers have lower precision and recall for the phrasal translation system, contradicting the results for the other two groups. A careful analysis of the data revealed that this experiment was spoiled by the three searchers that made the experiment remotely (see discussion below).

A detailed discussion of each of the three experiments follows.

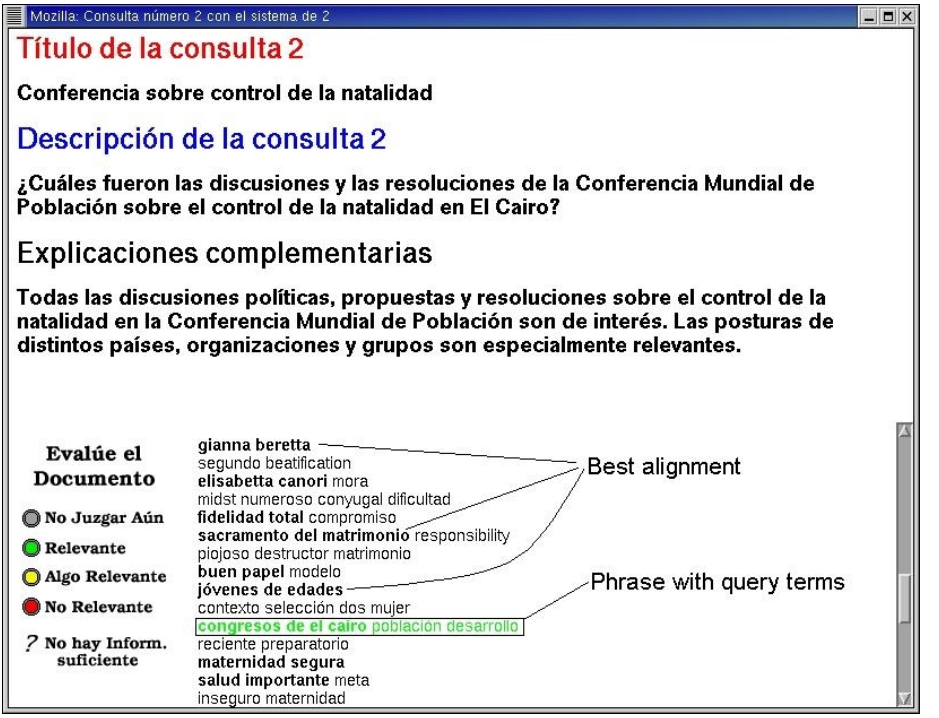

Fig. 2. Search interface: Phrases system

Table 1. Overview of results.

Main (Low level of English)

System	P	R	$F_{0.8}$	$F_{0.2}$
Systran MT	.48	.22	.28	.21
Phrases	.47(-2%)	.34(+52%)	.35(+25%)	.32(+52%)

Mid level of English

System	P	R	$F_{0.8}$	$F_{0.2}$
Systran MT	.62	.31	.41	.31
Phrases	.46(-25%)	.25(-19%)	.30(-26%)	.24(-22%)

High level of English

System	P	R	$F_{0.8}$	$F_{0.2}$
Systran MT	.58	.34	.42	.34
Phrases	.53(-12%)	.45(+32%)	.39(-7%)	.38(+11%)

4.1 Low Level of English Proficiency (Main Experiment)

The results of this experiment, detailed by searcher and topic, can be seen in Table 2. Looking at the average figures per searcher, the results are consistent

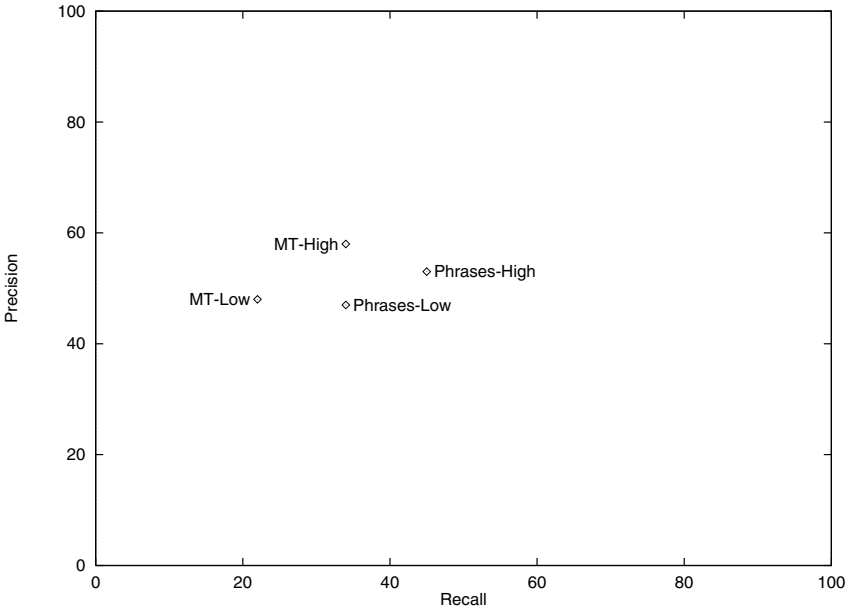

Fig. 3. High versus low English skills.

except for searcher 1 (with a very low recall) and searcher 5 (with very low recall and precision):

- Within this group, searcher 1 was the only one that carried out the experiment remotely, and problems with the net connection seriously affected recall for both systems and all topics. Unfortunately, this problem also affected three searchers in the mid-level English group and one in the high-level group.
- Examining the questionnaires filled in by Searcher 5, we concluded that he did not understand the task at all. He did not mark relevant documents in any of the questions, apparently judging the quality of the translations.

Only one of the eight searchers was familiar with MT systems, and most of them had little experience with search engines.

In the questionnaires, most searchers prefer the phrasal system, arguing that the information was more concise and thus decisions could be made faster. However, they felt that the phrases system demanded more interpretation from the user. The MT system was perceived as giving more detailed information, but too dense to reach easy judgments. All these impressions are coherent with the Precision/Recall figures obtained, and confirm our hypothesis about potential benefits of phrasal pseudo translations.

Table 2. Low Level of English (main experiment)

(Runs with the phrase system are in **boldface**, runs with MT in normal font)

Precision

User\Topic	T-1	T-2	T-3	T-4	Avg.
U-L-01	1	**0**	1	**0**	0.5
U-L-02	**1**	0.23	**0.66**	1	0.72
U-L-03	1	**0.34**	1	**0.25**	0.64
U-L-04	**1**	0.09	**0.33**	0	0.35
U-L-05	0	**0.2**	0	**0**	0.05
U-L-06	**1**	0	**0.57**	0.16	0.43
U-L-07	1	**0**	1	**0.33**	0.58
U-L-08	**0.95**	0.03	**1**	0.25	0.55
Avg.	0.86	0.11	0.69	0.24	0.47

Recall

User\Topic	T-1	T-2	T-3	T-4	Avg.
U-L-01	0.02	**0**	0.16	**0**	0.04
U-L-02	**0.19**	0.5	**1**	0.5	0.54
U-L-03	0.08	**0.93**	0.66	**0.5**	0.54
U-L-04	**0.11**	0.06	**0.5**	0	0.16
U-L-05	0	**0.18**	0	**0**	0.04
U-L-06	**0.13**	0	**0.66**	0.5	0.32
U-L-07	0.11	**0**	0.5	**0.5**	0.27
U-L-08	**0.55**	0.06	**0.33**	0.5	0.36
Avg.	0.14	0.21	0.47	0.31	0.28

$F_{0.2}$

User\Topic	T-1	T-2	T-3	T-4	Avg.
U-L-01	0.02	**0**	0.19	**0**	0.05
U-L-02	**0.22**	0.40	**0.90**	0.55	0.51
U-L-03	0.09	**0.69**	0.70	**0.41**	0.47
U-L-04	**0.13**	0.06	**0.45**	0	0.16
U-L-05	0	**0.18**	0	**0**	0.04
U-L-06	**0.15**	0	**0.63**	0.35	0.28
U-L-07	0.13	**0**	0.55	**0.45**	0.28
U-L-08	**0.60**	0.05	**0.38**	0.41	0.36
Avg.	0.16	0.17	0.47	0.27	0.26

$F_{0.8}$

User\Topic	T-1	T-2	T-3	T-4	Avg.
U-L-01	0.09	**0**	0.48	**0**	0.14
U-L-02	**0.53**	0.25	**0.70**	0.83	0.57
U-L-03	0.30	**0.38**	0.90	**0.27**	0.46
U-L-04	**0.38**	0.08	**0.35**	0	0.20
U-L-05	0	**0.19**	0	**0**	0.04
U-L-06	**0.42**	0	**0.58**	0.18	0.29
U-L-07	0.38	**0**	0.83	**0.35**	0.39
U-L-08	**0.82**	0.03	**0.71**	0.27	0.45
Avg.	0.36	0.11	0.56	0.23	0.31

4.2 Mid Level of English Proficiency

The results for this group (see Table 3) are apparently incompatible with the other two experiments. Taking a close look at the user averages, we detected that three users have extremely low recall figures, and these are precisely the users that performed the experiment remotely. Excluding them, the average recall would be similar for both systems. Of course the lesson learned from this spoiled experiment is that we have to be far more careful keeping the experiment conditions stable (and that we should not rely on Internet for this kind of experiment!).

4.3 High Level of English Proficiency

The detailed results for the group with good language skills can be seen in Table 4. Again, one searcher deviates from the rest with very low average recall, the only one that performed the experiment remotely (searcher 6). Aside from this, apparently better English skills lead to higher recall and precision rates. This is a reasonable result, as untranslated words can be understood, and translation errors can more easily be tracked back. Precision is 12% lower with the phrasal system, but recall is 32% higher. Overall, $F_{0.8}$ is higher for the MT system, and $F_{0.2}$ is higher for the phrasal system.

Table 3. Mid Level of English

(Runs with the phrase system are in **boldface**, runs with MT in normal font)

Precision

User\Topic	T-1	T-2	T-3	T-4	Avg.
U-M-01	1	**0**	1	**0**	0.5
U-M-02	**1**	0	**1**	0	0.5
U-M-03	1	**0.26**	1	**0.5**	0.69
U-M-04	**1**	0.31	**0**	0	0.32
U-M-05	1	**0.36**	1	**0.33**	0.67
U-M-06	**0.81**	0.30	**0.66**	0.33	0.52
U-M-07	1	**0**	1	**0**	0.5
U-M-08	**0.90**	0	**0.66**	1	0.64
Avg.	0.96	0.15	0.79	0.27	0.54

Recall

User\Topic	T-1	T-2	T-3	T-4	Avg.
U-M-01	0.11	**0**	0.66	**0**	0.19
U-M-02	**0.02**	0	**0.16**	0	0.04
U-M-03	0.13	**0.5**	0.5	**0.5**	0.40
U-M-04	**0.13**	0.31	**0**	0	0.11
U-M-05	0.11	**0.68**	0.66	**0.5**	0.48
U-M-06	**0.25**	0.87	**0.66**	0.5	0.57
U-M-07	0.08	**0**	0.16	**0**	0.06
U-M-08	**0.27**	0	**0.33**	1	0.4
Avg.	0.13	0.29	0.39	0.31	0.28

$F_{0.2}$

User\Topic	T-1	T-2	T-3	T-4	Avg.
U-M-01	0.13	**0**	0.70	**0**	0.20
U-M-02	**0.02**	0	**0.19**	0	0.05
U-M-03	0.15	**0.42**	0.55	**0.5**	0.40
U-M-04	**0.15**	0.31	**0**	0	0.11
U-M-05	0.13	**0.57**	0.70	**0.45**	0.46
U-M-06	**0.29**	0.63	**0.66**	0.45	0.50
U-M-07	0.09	**0**	0.19	**0**	0.07
U-M-08	**0.31**	0	**0.36**	1	0.41
Avg.	0.15	0.24	0.41	0.3	0.27

$F_{0.8}$

User\Topic	T-1	T-2	T-3	T-4	Avg.
U-M-01	0.38	**0**	0.90	**0**	0.32
U-M-02	**0.09**	0	**0.48**	0	0.14
U-M-03	0.42	**0.28**	0.83	**0.5**	0.50
U-M-04	**0.42**	0.31	**0**	0	0.18
U-M-05	0.38	**0.39**	0.90	**0.35**	0.50
U-M-06	**0.55**	0.34	**0.66**	0.35	0.47
U-M-07	0.30	**0**	0.48	**0**	0.19
U-M-08	**0.61**	0	**0.55**	1	0.54
Avg.	0.39	0.16	0.6	0.27	0.35

Besides having better English skills, searchers had more experience using graphical interfaces, search engines and Machine Translation programs. In agreement with the first group, they felt that the MT system gave too much information, and they also complained about the quality of the translations. However, overall they preferred the MT system to the phrasal one: translated phrases permitted faster judgments, but the searcher needed to add more subjective interpretation of the information presented. All these subjective impressions are in agreement with the final precision/recall figures.

5 Conclusions

Although the number of searchers does not allow for clear-cut conclusions, the results of the evaluation indicate that summarized translations, and in particular phrasal equivalents in the searcher's language, might be more appropriate for document selection than full-fledged MT. Our purpose is to reproduce a similar experiment with more users, and better-controlled experimental conditions, to have a better testing of our hypothesis in a near future.

As a side conclusion, we have proved that phrase detection and handling with shallow NLP techniques is feasible for large-scale IR collections. The major

Table 4. High Level of English

(Runs with the phrase system are in **boldface**, runs with MT in normal font)

Precision

User\Topic	T-1	T-2	T-3	T-4	Avg.
U-H-01	1	**0.27**	1	**0**	0.56
U-H-02	**0.91**	0.35	**0.8**	0.33	0.59
U-H-03	0.83	**0.17**	1	**0.66**	0.66
U-H-04	**1**	0	**1**	0.5	0.62
U-H-05	1	**0.34**	0.83	**0.25**	0.60
U-H-06	**0**	0.33	**0.66**	0	0.24
U-H-07	1	**0.33**	1	**0.13**	0.61
U-H-08	**1**	0.21	**1**	0	0.55
Avg.	0.84	0.25	0.91	0.23	0.55

Recall

User\Topic	T-1	T-2	T-3	T-4	Avg.
U-H-01	0.05	**0.37**	0.66	**0**	0.27
U-H-02	**0.30**	0.93	**0.66**	0.5	0.59
U-H-03	0.13	**0.18**	0.5	**1**	0.45
U-H-04	**0.02**	0	**0.83**	0.5	0.33
U-H-05	0.30	**1**	0.83	**0.5**	0.65
U-H-06	**0**	0.25	**0.33**	0	0.14
U-H-07	0.16	**0.62**	0.33	**1**	0.52
U-H-08	**0.08**	0.43	**0.33**	0	0.21
Avg.	0.13	0.47	0.55	0.43	0.39

$F_{0.2}$

User\Topic	T-1	T-2	T-3	T-4	Avg.
U-H-01	0.06	**0.34**	0.70	**0**	0.27
U-H-02	**0.34**	0.69	**0.68**	0.45	0.54
U-H-03	0.15	**0.17**	0.55	**0.90**	0.44
U-H-04	**0.02**	0	**0.85**	0.5	0.34
U-H-05	0.34	**0.72**	0.83	**0.41**	0.57
U-H-06	**0**	0.26	**0.36**	0	0.15
U-H-07	0.19	**0.52**	0.38	**0.42**	0.37
U-H-08	**0.09**	0.35	**0.38**	0	0.20
Avg.	0.14	0.38	0.59	0.33	0.36

$F_{0.8}$

User\Topic	T-1	T-2	T-3	T-4	Avg.
U-H-01	0.20	**0.28**	0.90	**0**	0.34
U-H-02	**0.64**	0.39	**0.76**	0.35	0.53
U-H-03	0.39	**0.17**	0.83	**0.70**	0.52
U-H-04	**0.09**	0	**0.96**	0.5	0.38
U-H-05	0.68	**0.39**	0.83	**0.27**	0.54
U-H-06	**0**	0.31	**0.55**	0	0.21
U-H-07	0.48	**0.36**	0.71	**0.15**	0.42
U-H-08	**0.30**	0.23	**0.71**	0	0.31
Avg.	0.34	0.26	0.78	0.24	0.40

bottleneck, Part-Of-Speech tagging, can be overcome with heuristic simplifications that do not compromise the usability of the results, at least in the present application.

Acknowledgments. This work has been funded by the Spanish *Comisión Interministerial de Ciencia y Tecnología*, project *Hermes* (TIC2000-0335-C03-01).

References

1. J. Carmona, S. Cervell, L. Màrquez, M. A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. An environment for morphosyntactic processing of unrestricted spanish text. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*, 1998.
2. Douglas W. Oard and Julio Gonzalo. The CLEF 2001 interactive track. In Carol Peters, editor, *Proceedings of CLEF 2001*, Lecture Notes for Computer Science, Springer, forthcoming.
3. 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.
4. Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, 1994.