NE Recognition and Keyword-based Relevance Feedback in Mono and CL Automatic Speech Transcriptions Retrieval

F. López-Ostenero, V. Peinado, V. Sama & F. Verdejo

CLEF 2005. Vienna, Austria.
September 22th 2005

Dpto. de Lenguajes
y Sistemas Informáticos
E.T.S.I. Informática, UNED

C/ Juan del Rosal, 16
Ciudad Universitaria
28040 Madrid

(+34) 91 398 7983
(+34) 91 398 6535

Goals

- Test the suitability of our translation resources for a new track
- Test different strategies to clean the automatic transcriptions
- Check if Proper Noun Recognition can help to improve retrieval over automatic speech transcriptions
- Compare the efectiveness of manual and automatic keywords in a keyword-based pseudo-relevance feedback approach

Cleaning strategies

Starting poing: ASR2004A field

Join all the occurrences of a list of single characters:
l i e b b a c h a r d → liebbachard
Remove all extra occurences of duplicated words:
"yes yes yes yes" → "yes"

→ CLEAN collection

Split documents of CLEAN collection in character 3-grams

→ 3GRAMS collection

Perform a morphological analysis over CLEAN collection and remove all words that cannot act as noun, adjective or verb

→ MORPHO collection

Perform a full Part of Speech tagging over CLEAN collection and delete all words except nouns, adjectives and verbs

→ POS collection

Submitted runs

Ranking	MAP	Run	Language
20	0.0934	mono-pos	English
21	0.0918	mono-morpho	English
29	0.0706	mono-3grams	English
32	0.0373	trans-pos	Spanish
33	0.0370	trans-morpho	Spanish

Runs based on a query translation approach (using Pirkola's structured queries) and INQUERY retrieval engine

•Scores far from best monolingual and Spanish crosslingual runs: room for improvement
•MAP scores of **morpho** and **pos** very similar, what's the influence of cleaning strategies?
•Character 3-grams scores worse than full word retrieval

Only 5 different runs: not enough data to obtain clear conclusions

Proper noun identification

We used a shallow entity recognizer to identify proper nouns in topics:

- Monolingual: we identify proper nouns in English topics and we structure the query tagging them with a proximity operator

- Crosslingual: we also identify proper nouns in Spanish topics but, which proper nouns should be translated and which ones should not?

  a) if a proper noun appears in the SUMMARY field of documents, we assume that it should not be translated and we tag it using a proximity operator
  b) otherwise we try to translate the proper noun

  La historia de *Varian Fry* y el *Comité de Rescates de Emergencia* …

  don't translate — try to translate

Pseudo Relevance Feedback

Five collections to study a keyword-based pseudo relevance feedback:

1) AUTOKEYWORD2004A1 field (to build up AK1 collection)
2) AUTOKEYWORD2004A2 field (to build up AK2 collection)
3) Mix of 1 and 2
   a) single field keyword score, according to the position: 1st keyword = 20; 2nd keyword = 19 ...
   b) if a keyword appears in both fields, its final score is the sum of both single field scores
   c) select the top 20 scored keywords (to build up AK12 collection)
4) MANUALKEYWORD field (to build up MK collection)
5) Mix of 3 and 4
   a) Select the *n* manual keywords from 4
   b) If $n < 20$, add the $20 - n$ first keywords from 3 (to build up MKAK12 collection)

Pseudo relevance feedback procedure:

a) Launch a plain query (without keyword expansion)
b) Retrieve keywords from top 10 retrieved documents
c) Mix these keywords using algorithm described to create AK12 collection
d) Expand query using top 20 keywords

Combining techniques

$$\begin{pmatrix} \text{mono} \\ \text{trans} \end{pmatrix} \times \begin{pmatrix} \text{noent} \\ \text{ent} \end{pmatrix} \times \begin{pmatrix} \text{3grams} \\ \text{clean} \\ \text{morpho} \\ \text{pos} \end{pmatrix} \times \begin{pmatrix} \text{NO} \\ \text{AK1} \\ \text{AK2} \\ \text{AK12} \\ \text{MK} \\ \text{MKAK12} \end{pmatrix}$$

language *proper noun* *cleaning method* *relevance feedback*

Total combinations: 2 x 2 x 4 x 6 = 96 runs

No 3grams in crosslingual runs:

- mono: 2 x 4 x 6 = 48 runs
- trans: 2 x 3 x 6 = 36 runs
- real number or runs: 48 + 36 = 84 runs

Results

Ranking	MAP	Run	Variation wrt UNED submitted runs
1	0.2595	mono-ent-morpho-MK	277.8% (monolingual)
13	0.2036	trans-ent-pos-MK	545.8% (crosslingual)
31	0.0934	mono-noent-pos-NO	100% (monolingual)
63	0.0706	mono-noent-3grams-NO	75.6% (monolingual)
73	0.0373	trans-noent-pos-NO	100% (crosslingual)

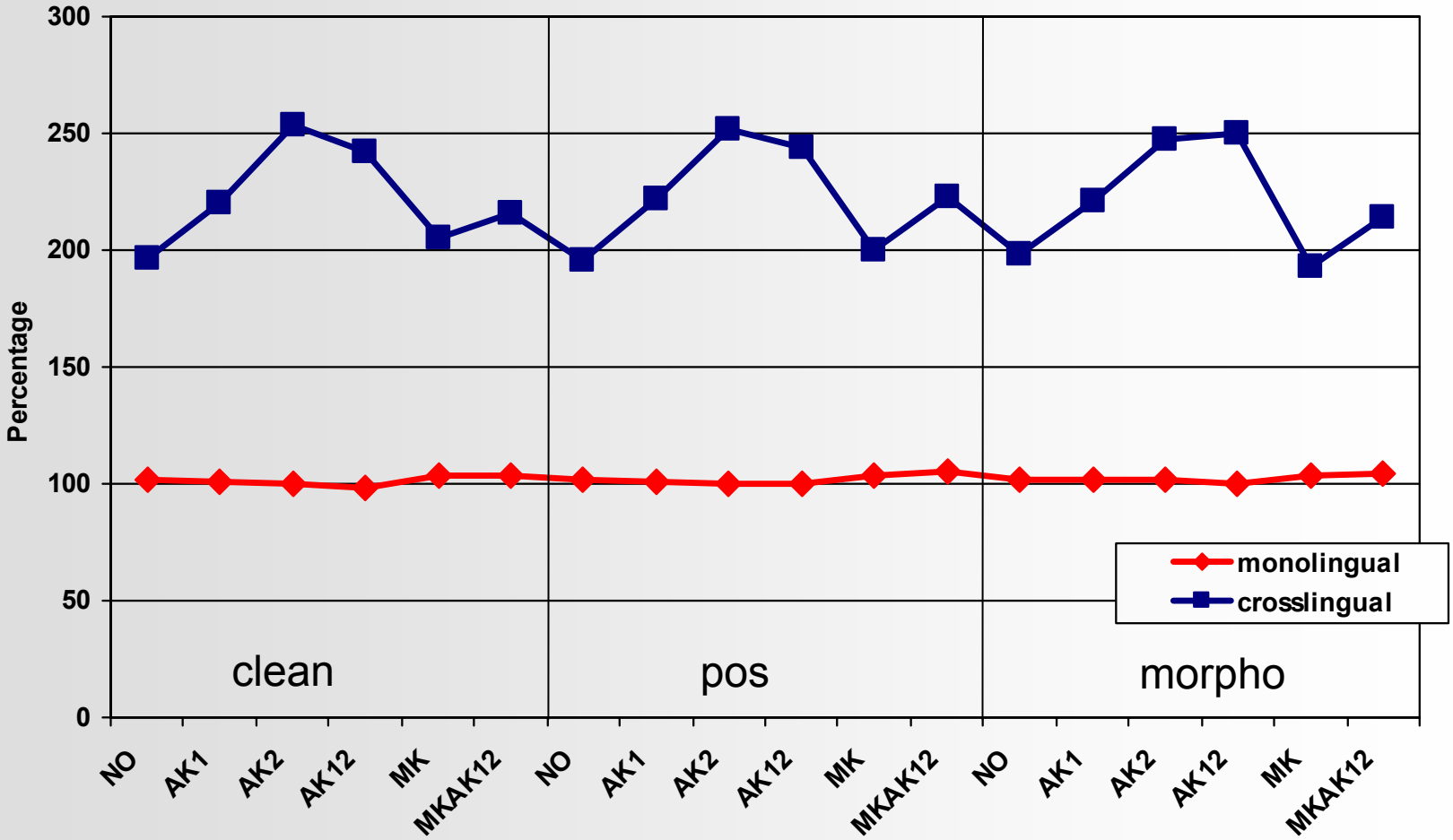
Preliminary conclusions:

- Monolingual improvement: 277.8%
- Crosslingual improvement: 545.8%
- Best strategies:
  - PRF using MK or MKAK12 collections
  - Use proper noun recognition
- Monolingual 3-grams scores poorly, reaching a 27.2% of our best run

Influence of proper nouns

Each point represents MAP ent / MAP noent in percentage

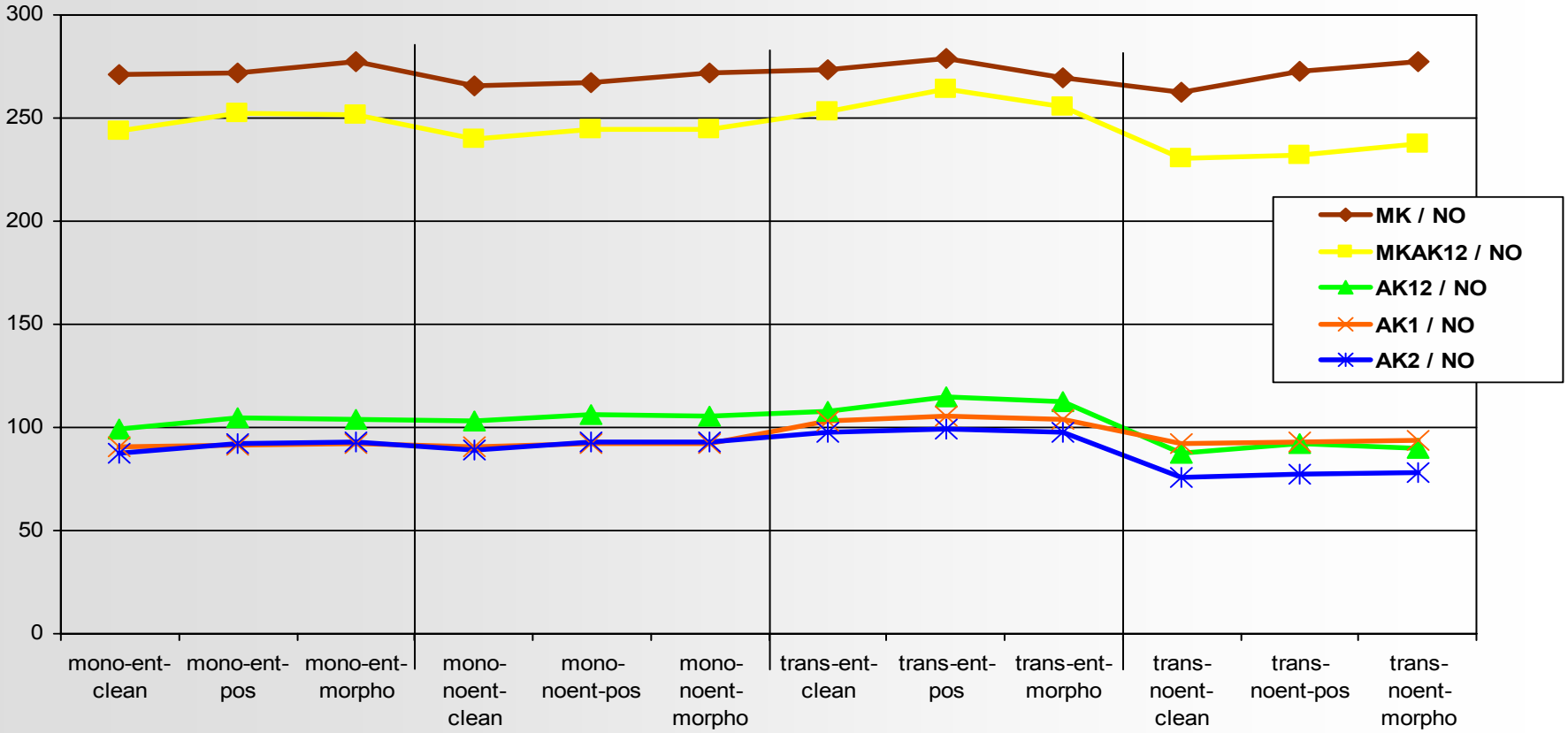
- monolingual: increment worthless and probably statistically not relevant
- crosslingual: proper noun detection increment MAP more than twice



Influence of relevance feedback

Each point represents MAP (rf method) / MAP (NO rf) in percentage

- MK: the best option, but when combined with AK12 MAP decreases
- AK12 usually better than AK1 or AK2. More stable in monolingual, better in crosslingual with entities and worse in crosslingual without entities
- AK1 and AK2 identical in monolingual, AK1 better in crosslingual



Conclusions and future work

- **The use of a shallow entity recognizer to identify proper nouns seems to be very useful, specially in a crosslingual environment where MAP increases 221,9% on average**

- Cleaning methods based on full words (clean, morpho and pos) show no significative differences, but character 3-grams approach seems to be not useful for this task

- Pseudo Relevance Feedback using manually generated keywords shows to be the best option to improve the performance of the retrieval, with an average of 271.6% MAP regarding no relevance feedback

- Perform further analysis over the results, including statistical relevance tests

- Try a different approach to identify proper nouns in the automatic transcriptions or in the automatic keyword fields, instead of using the manual summary of the transcriptions

NE Recognition and Keyword-based Relevance Feedback in Mono and CL Automatic Speech Transcriptions Retrieval

F. López-Ostenero, V. Peinado, V. Sama & F. Verdejo

CLEF 2005. Vienna, Austria.
September 22th 2005

Dpto. de Lenguajes
y Sistemas Informáticos
E.T.S.I. Informática, UNED

C/ Juan del Rosal, 16
Ciudad Universitaria
28040 Madrid

(+34) 91 398 7983
(+34) 91 398 6535