UNED bilingual experiments at WebCLEF

Cross Language Evaluation Forum 2005

Javier Artiles, Víctor Peinado, Anselmo Peñas, Felisa Verdejo

NLP & IR Group
http://nlp.uned.es
Dept. de Lenguajes y Sistemas Informáticos
ETSI de Informática - UNED

Vienna. September 23, 2005

UNED's participation at WebCLEF 2005

- Indexing EuroGov
- Translating topics
- Experiments
- Results

Problems

- Monolingual task
- Bilingual EN-ES task
- Lessons

Indexing EuroGov

Bilingual EN→ES task: English topics with target webpages written in Spanish.

- Lucene's API.
- One index containing ES + UNKNOWN pages.
- Indexed by fields: title, metadata, headers, body. . .
- No stemming or lemmatization.
- Tokenization with Lucene's `StandardAnalyzer`.

Do we have to translate or not?

Given a word w and an index $Lang$, we decide whether to translate w or not, by computing the word relative frequency wrt the number of documents in the corresponding index:

$$F(w)_{Lang} = \frac{tf(w)}{N_{Lang}} \tag{1}$$

- In the bilingual task, if $F_{w_{ES}} > F_{w_{EN}}$ then we assume w must remain untranslated.

Description of Moncloa Palace

- Otherwise, we attempt to lemmatize and translate it with our dictionary (Vox + EuroWordNet + FreeDict).

Translating OOV words

A Spanish document containing the English word w_{EN} might also contain its translation into Spanish e.g. [Vogel et al., SIGIR05]

Given an English OOV word w_{EN}:

1. We google for Spanish pages containing w_{EN}.
2. We take the 10 most frequent Spanish words (after removing stopwords) from the 40 first snippets retrieved by Google → candidate translations
3. We search for English pages containing each candidate and count up every time it co-occurs with w_{EN} in the snippets → inverse translation
4. We rank the candidates and choose the most frequent in 2) and 3) as the ultimate translation.

Translating OOV words

Strategy for the preservation of the Cantabrian brown bear

1. We google for Spanish pages containing *Cantabrian*.
2. We take the 10 most frequent words appearing in the snippets:

   cantabria (16), spain (14), diccionarios (10), mountains (9),
   glosarios (6), términos (6), turismo (5), región (5), sea (5)

3. We search for English pages containing each candidate and count up every time it co-occurs with *Cantabrian*.
4. We rank the candidates and choose *cantabria* as the ultimate translation.

Experiments

Terms are combined with the `AND` operator in a full boolean query.

- Baseline: search over the body field and let the search engine rank the results.
- Proposal: order the fields (title, metadata, headings, body).
  1. Launch the query over title fields.
  2. If we don't get 50 pages, launch the query over metadata fields and append the results, removing duplicates, if any.
  3. If we don't have 50 pages, launch the query over heading fields and append the results, removing duplicates, if any.
  4. . . .

If we cannot reach 50 results yet, we repeat the process using `OR`.

Results

	baseline	proposal	variation
Avg success at 1	0.02	0.08	+300%
Avg success at 5	0.07	0.10	+43%
Avg success at 10	0.10	0.12	+20%
Avg success at 20	0.17	0.13	-23%
Avg success at 50	0.26	0.21	-19%
MRR over 134 topics	0.05	0.09	+80%

To favor searches over the most descriptive fields seems promising.

Problems in the Monolingual task

Initial idea	Use language identification both for topics and documents and build one index per language.
Problem	Language identification was very noisy.
Solution	No language identification, build a unique index and launch queries.

The index became a monster and our system was too slow.
We didn't have time to tame Lucene.
For the same reasons, we couldn't try the Multilingual task.

Problems in the Bilingual EN-ES task

Not enough time to implement and test most of our initial ideas, such as:

- Use text anchors pointing to a page as the page's descriptors.
- Evaluate the impact of our OOV translation process.
- Improve the OOV translation process, refining the way we select the translation among all candidates.

Lessons we've learned

- Don't sign up in too many CLEF tracks.
- Dealing with huge amount of data is not straightforward.
- Don't trust blindly language identification.
- Be careful with the translations: 1/3 of the Spanish topics were incorrectly translated into English.

The slide you were waiting for...

Questions?

Thank you!