COLESIR at CLEF 2007:

from English to French via Character *N*-Grams

Jesús Vilares

Computer Science Dept.

University of A Coruña

jvilares@udc.es

Michael P. Oakes

School of Computing and Technology

University of Sunderland

Michael.Oakes@sunderland.ac.uk

Manuel Vilares

Computer Science Dept.

University of Vigo

vilares@uvigo.es

Index

- Introduction
- Previous approaches
- Our proposal
- Evaluation
- Conclusions and future work

Index

- **Introduction**
- Previous approaches
- Our proposal
- Evaluation
- Conclusions and future work

Translation in CLIR

- Techniques of **Machine Translation (MT)**
  - Softened restrictions
    - Not limited to just one translation
    - Not limited by syntax
- ***Conventional* MT tools** (e.g., SYSTRAN)
  - Single well-formed translation
  - Dismisses advantages of MT in CLIR

Translation in CLIR (cont.)

- **Bilingual dictionaries**
  - Problems with out-of-vocabulary words (misspellings, unknown words)
  - Normalization
  - Word-Sense Disambiguation (WSD)
- **Parallel corpora**
  - Automatic generation of dictionaries:
    - Collocations
    - Association measures
  - Probabilistic translation measure
  - No normalization

Character *N*-Grams

$$\texttt{tomatoes} \xrightarrow{n=5} \{\ \texttt{-tomat-}\ ,\ \texttt{-omato-}\ ,\ \texttt{-matoe-},\ \texttt{-atoes-}\ \}$$

Applications:

- Language recognition
- Misspelling processing
- Information Retrieval
  - Reduction of vocabulary size (dictionary)
  - Asian languages (no delimiters)

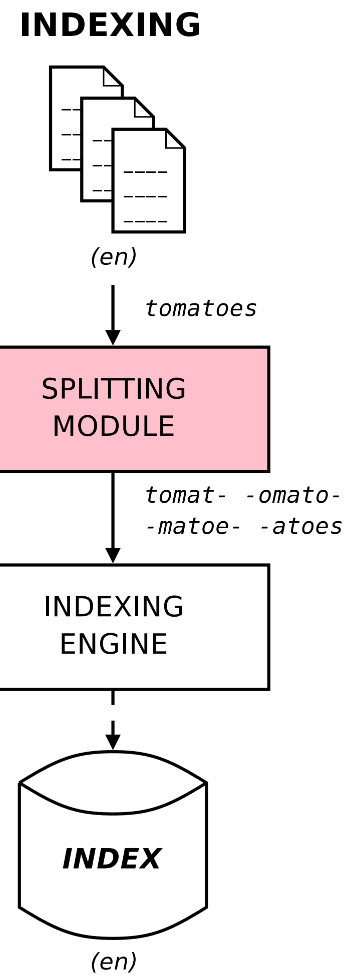
Index

- Introduction
- **Previous approaches**
- Our proposal
- Evaluation
- Conclusions and future work

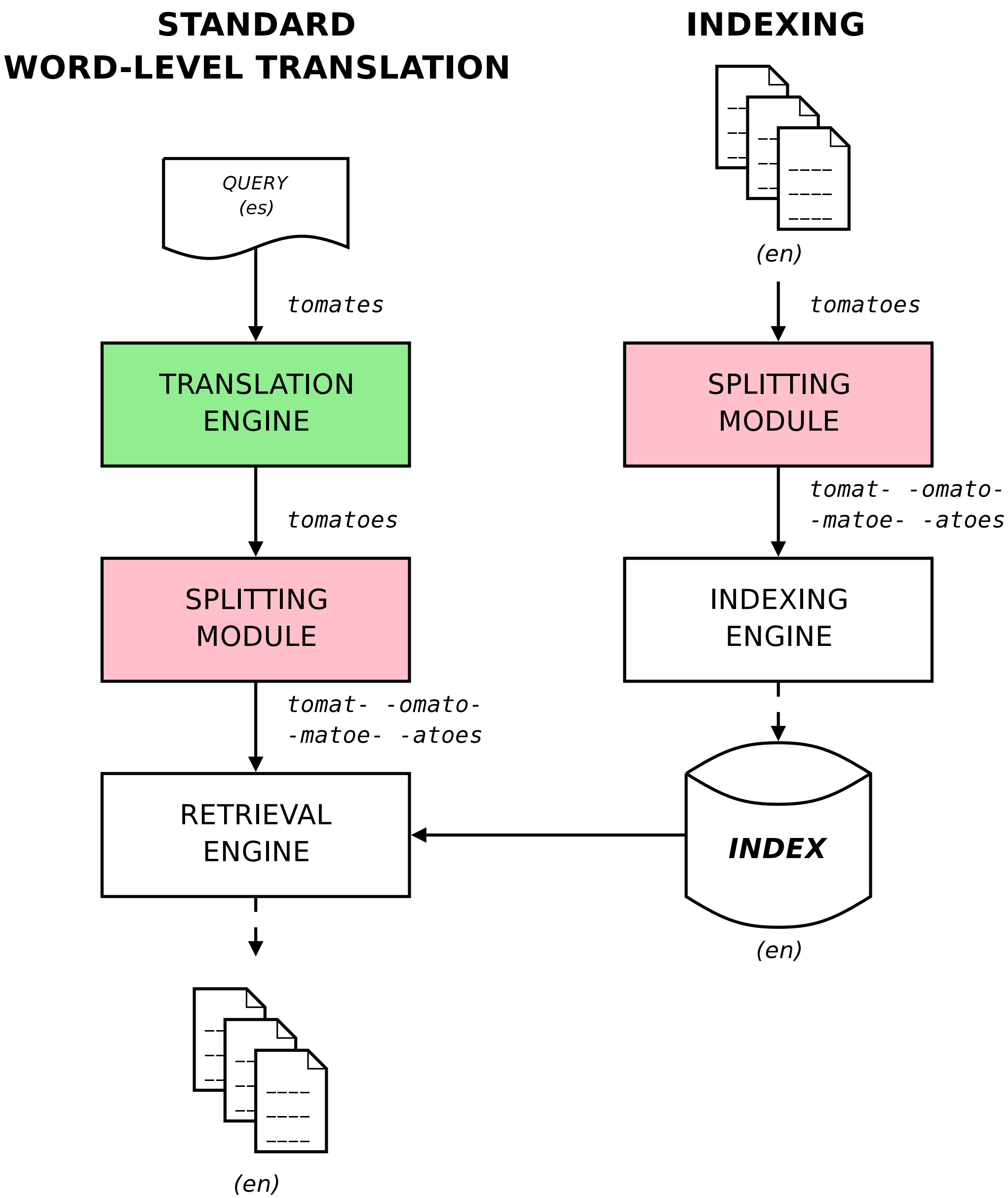
McNamee and Mayfield, 2004

- **No word normalization**
- **Language-independent**:
  - No language-specific processing
  - Applicable to very different languages
- **Knowledge-light approach**:
  - Minimal linguistic information and resources
- **Robustness**:
  - Out-of-vocabulary words

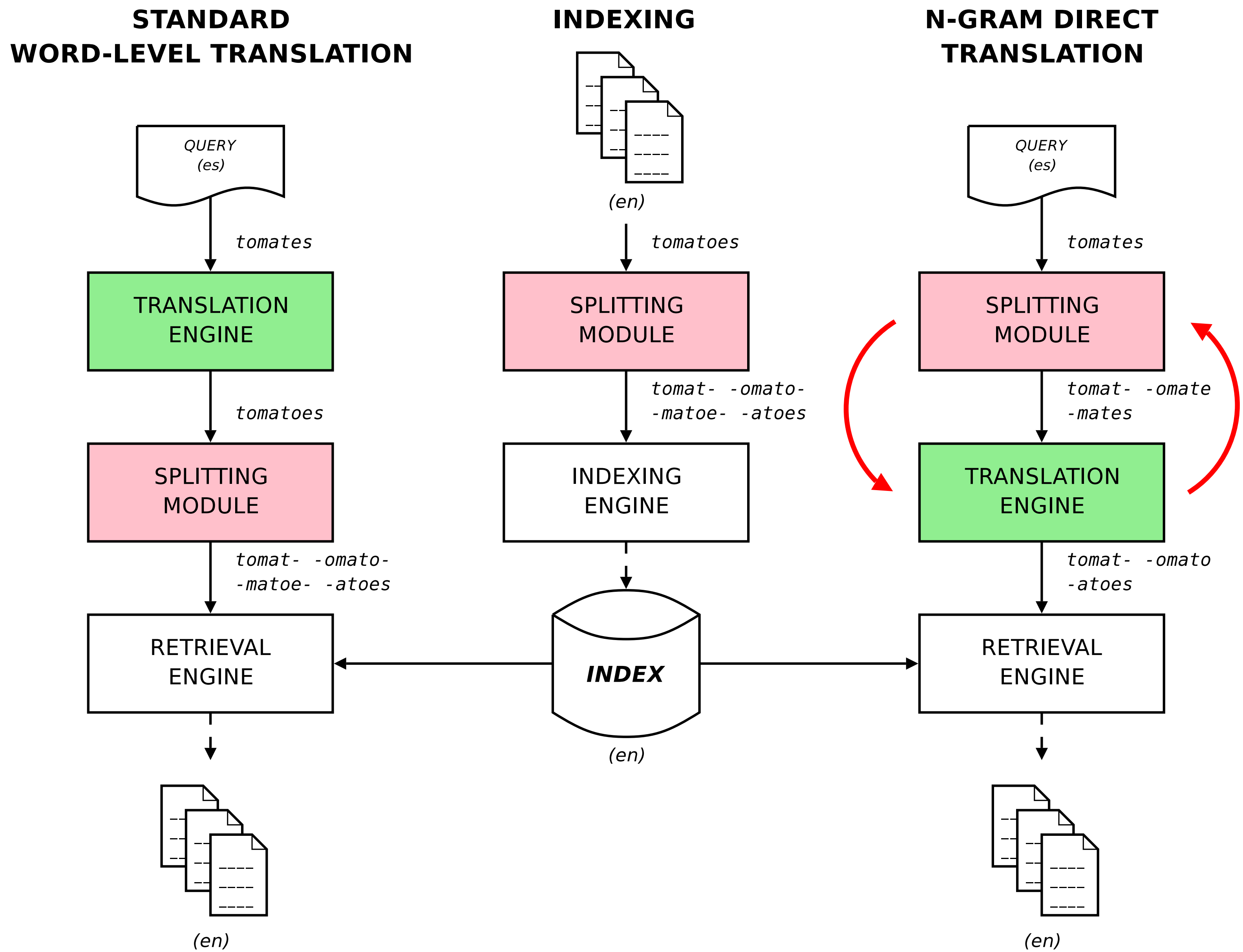
McNamee and Mayfield, 2004 (cont.)



McNamee and Mayfield, 2004 (cont.)



McNamee and Mayfield, 2004 (cont.)



N-Gram Alignment Algorithm

- **Input:** parallel corpus aligned at paragraph-level
  - Text splitted into *n*-grams

- **Process**: for each source *n*-gram of source language:
  1. To locate source language paragraphs containing it
  2. To identify parallel paragraphs in target language
  3. To calculate **translation score** for each *n*-gram in target paragraphs (**ad-hoc association measure**).
  4. **Potential translation**: target *n*-gram with highest score.

- **Output:** *n*-gram-level alignment

N-Gram Alignment Algorithm (cont.)

- **Drawbacks:**
  - **Very slow** (several days): **not accurate for testing**
  - Single translation

Index

- Introduction
- Previous approaches
- **Our proposal**
- Evaluation
- Conclusions and future work

Goals

- **Testing tool**
- **To speed up the training process**
- Multiple translations
- Freely available resources
  - More transparency
  - Reduce effort

Differences

- **Freely available resources**:
  - Parallel corpus: EUROPARL *(Koehn, 2005)*
  - Statistical aligner: GIZA++ *(Och and Ney, 2003)*
  - Retrieval engine: TERRIER *(`http://ir.dcs.gla.ac.uk/terrier/`)*
- **Standard association measures**:
  - Dice coefficient
  - Mutual Information
  - Log-likelihood
- **Alignment in two phases**:
  1. Word-level alignment
  2. *N*-gram-level alignment

N-Gram Alignment Algorithm

- **Input:** parallel corpus aligned at paragraph-level
- **Process**: two phases
  1. **Word-level alignment** using GIZA++ (slowest): **filtering**
  2. ***N*-gram-level alignment**:
     - Aligned words as **weighted word-level parallel corpus**
     - **Association measures** between cooccurring *n*-grams
       - **Likelihood of cooccurrences weighted according to their alignment probabilities (from word-level alignment)**
- **Output:** *n*-gram-level alignment

N-Gram Alignment Algorithm (cont.)

Optimizations:

- **Input word-translation probability threshold *W*** (W=0.15)
  - Input word pairs / output *n*-gram pairs: $\sim$95 % reduction
- **Bidirectional word alignment** (*EN2FR* $\cap$ *FR2EN*)
  - Input word pairs / output *n*-gram pairs: $\sim$50 % reduction

Index

Evaluation

- **English-to-French** run (*EN2FR*)
- 4-grams *(McNamee and Mayfield, 2004)*
- TERRIER retrieval engine: DFR paradigm
  - `InL2` weight
- **Corpus:** CLEF 2007 robust track *(Cross-Language Evaluation Forum)*

collection (FR)	***size***	***#docs.***	***#topics (EN)***
LeMonde 94 + SDA 94	243 MB	87,191	100 (*training*) 100 (*test*)

Querying

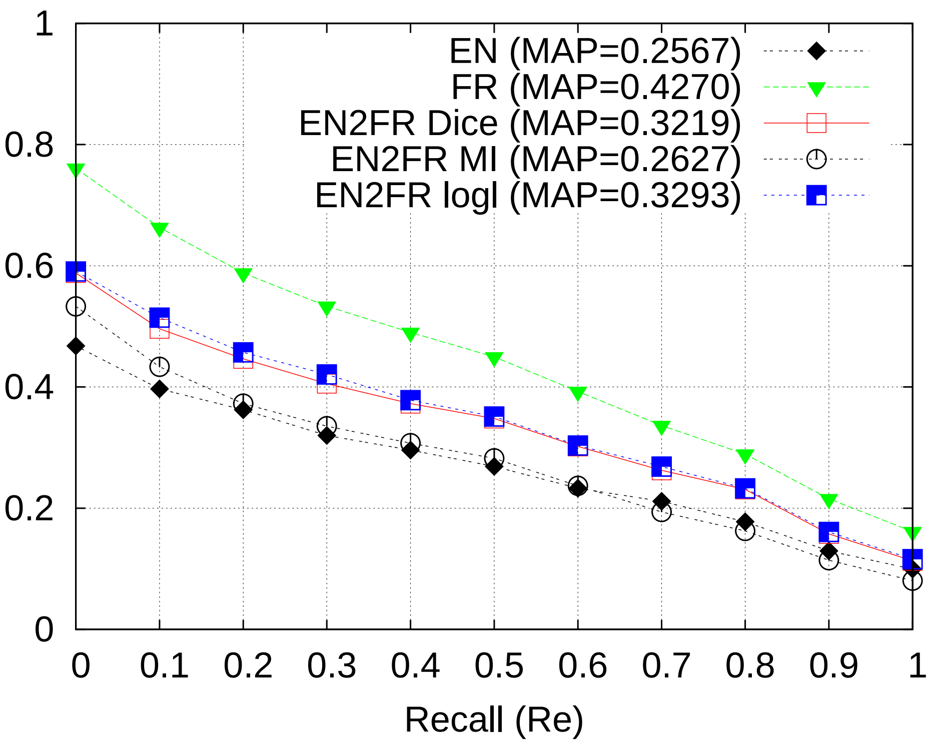
- `title` + `description` topic fields

- **Querying process:**
  - Split source language query into *n*-grams
  - Replaced by their N highest scored aligned target n-grams:
    - Tuned using **English-to-Spanish** experiments (*EN2ES*)

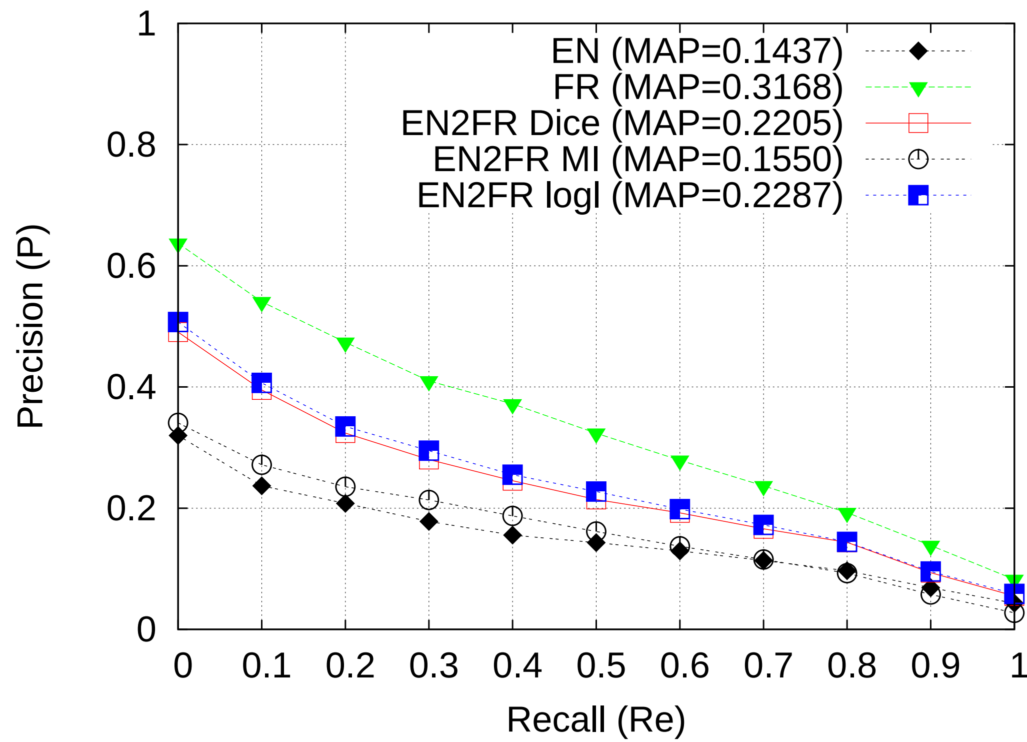
Dice coefficient	N=1
Mutual Information	N=10
Log-likelihood	N=1

  - Submit translated query

Precision vs. Recall

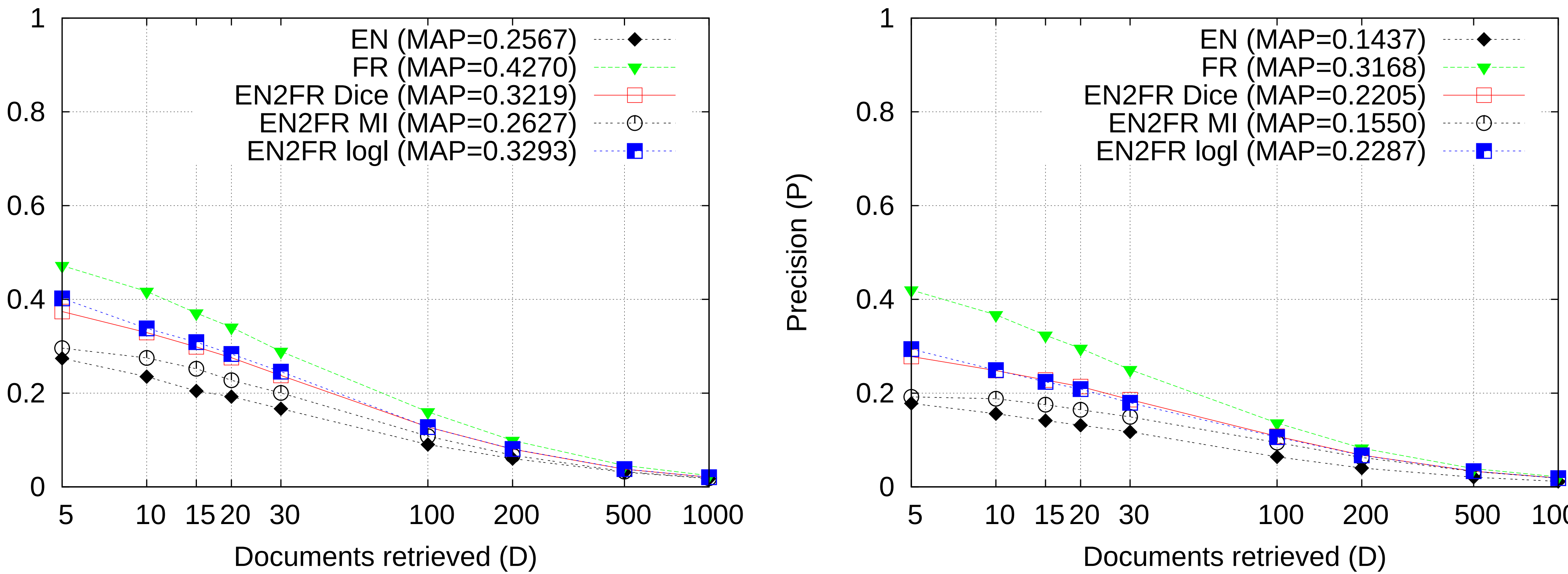


TRAINING set



TEST set

Precision at top *D* documents



TRAINING set *TEST set*

Index

- Introduction
- Previous approaches
- Our proposal
- Evaluation
- **Conclusions and future work**

Conclusions

- CLIR using ***n*-grams as indexing and translation units**
- ***N*-gram alignment in two phases: speeds up process**
  1. Word-level alignment (**concentrates complexity**)
  2. *N*-gram-level alignment
- **Optimizations** during **word-level alignment**:
  - Word-translation probability threshold
  - Bidirectional alignment
- **Dice and log-likelihood** perform better

Future work

- New languages
- Remove diacritics
- Remove stopwords and/or *stopngrams* (obtained automatically)
- Simplify word-level alignment (**bottleneck**)
- **Direct evaluation of *n*-gram alignments**

The End

www.grupocole.org

Go back to the beginning of the presentation

N-Gram Contingency Table

(U,V): (u, v)?

	V=v	V≠v	
U=u	O_{11}	O_{12}	$= R_1$
U≠u	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

N-Gram Contingency Table (cont.)

The likelihood of a cooccurrence is inherited from the probability of its containing word alignment:

$$P(ngram_{iu} \rightarrow ngram_{jv}) = P(word_u \rightarrow word_v)$$

$$\begin{array}{ccc}
\texttt{tomate} & \texttt{tomato} & 0.80 \\
\downarrow & \downarrow & \downarrow \\
\texttt{tomat- -omate} & \texttt{tomat- -omato} & 0.80 \\
\downarrow & \downarrow & \downarrow \\
\texttt{tomat-} & \texttt{tomat-} & 0.80 \\
\texttt{tomat-} & \texttt{-omato} & 0.80 \\
\texttt{-omate} & \texttt{tomat-} & 0.80 \\
\texttt{-omate} & \texttt{-omato} & 0.80
\end{array}$$

N-Gram Contingency Table (cont.)

Also reflected in the **contingency table**. E.g.:

$$O_{11}(ngram_{iu}, ngram_{jv}) = \sum_{\substack{uk/ngram_{iu} \in N(word_{uk}) \\ vk/ngram_{jv} \in N(word_{vk})}} P(word_{uk} \rightarrow word_{vk})$$

<u>tomat</u>e	<u>tomat</u>o	0.80
...	...	...
<u>tomat</u>itos	<u>tomat</u>oes	0.65
$\downarrow$	$\downarrow$	$\downarrow$
tomat-	tomat-	1.45 = 0.80+0.65

N-Gram Association Measures

- **Dice Coefficient:**

$$Dice(g_s, g_t) = log\frac{2O_{11}}{R_1 + C_1}$$

- **Mutual Information:**

$$MI(g_s, g_t) = log\frac{NO_{11}}{R_1 C_1}$$

- **Log-likelihood:**

$$logl(g_s, g_t) = 2\sum_{i,j} O_{ij}\ log\frac{NO_{ij}}{R_i C_j}\ .$$