Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR

Xabier Saralegi   Maddalen López de Lacalle

R&D
Elhuyar Foundation

7th international conference on Language Resources and Evaluation
LREC 2010, Valletta, Malta
2010/05/20

Outline

Introduction: Motivation

- CLIR = IR + language barrier
- Most CLIR technology based on Machine Translation Systems (MTS) or Parallel Corpora (PC)
  - MTS and PC resources **expensive or scarce** for most pair of languages, **specially for small languages**
- **Bilingual dictionaries** easier to obtain

Introduction: Bilingual Dictionaries

- Problems: **Translation ambiguity**, Out-of-Vocabulary words, Multi Word Expressions

Example

Query 80:

- *EU: “G7 gailurrean Napolin Errusiak jokatutako* **papera***“*
- *EN: “role played by Russia in the G7 summit in Naples in 1994”*
- **papera** *: paper, role. . .*

Introduction: Bilingual Dictionaries

- Problems: Translation ambiguity, **Out-of-Vocabulary words**, Multi Word Expressions

Example

Query 46:

- *EU: "Irakeko* **bahitura** *"*
- *EN: "***Embargo** *on Iraq"*

Introduction: Bilingual Dictionaries

- Problems: Translation ambiguity, Out-of-Vocabulary words, **Multi Word Expressions**

Example

Query 47:

- *EU: "Errusiarren* **esku hartzea** *Txetxenian"*
- *EN: "Russian* **intervention** *in Chechnya"*

Introduction: Objectives

- Objetives of this work
  - To analyse how each problem affects retrieval performance of a dictionary-based Basque-English CLIR system
  - To evaluate methods not based on parallel corpora to treat those problems

Outline

1. Introduction
2. Related work
   - Different Strategies
   - CLIR Frameworks based on query translation
3. Proposed query translation method
   - Experimental setup
   - Treatment of OOV words
   - MWE
   - Translation Selection
4. Evaluation
5. Conclusions

Different Strategies

- Translate $\longrightarrow$ collection or queries?
  - **Collection** $\longrightarrow$ richer context for translation selection (Oard, 1998)
  - **Query** $\longrightarrow$ most studied because it is more scalable (Hull and Grefenstette, 1996)
  - Best results: **Translating both**, merging corresponding rankings (McCarley, 1999)(Wang and Oard, 2003)

Outline

CLIR Frameworks based on query translation

(a) Post-translation Relevance Model (PTRM)

- The query is translated **independientely** and then a relevance model is used
- Query terms translated with PC or dict.
  - PC solves translation selection
  - Dict.: co-occurrence based method for solving selection (Monz and Dorr, 2005) (Gao et al., 2002)

CLIR Frameworks based on query translation

(b) Cross-lingual probabilistic relevance models (CLPRM)

- Translation process included in relevance model
- Query terms translated by PC or dict.
- All candidates are treated as a single token (Pirkola, 1998), or pondered with weights mined from PC (Darwish and Oard, 2003) or comparable corpora (Saralegi and Lopez de Lacalle, 2010)

$$TF_j(s_i) = \sum_{\{k|D_k \in T(s_i)\}} TF_j(D_k)$$

$$DF(Q_i) = |\cup_{\{k|D_k \in T(Q_i)\}} \{d|D_k \in d\}|$$

CLIR Frameworks based on query translation

(c) Cross-lingual language models (CLLM)

- Translation process included in relevance model
- Query terms translations PC or dict.
- Translation probabilities are included in a probabilistic model (Xu, Weischedel, and Nguyen, 2001)

$$P(Q_s|D_t) = \prod_{w \in Q_s} (((1-\lambda)P(w|G_s)) + \lambda(\sum_{t \in D_t} P(t|D_t)P(w|t)))$$

CLIR Frameworks based on query translation

- CLLM (c) better than CLPRM (b) when PC provided (Xu, Weischedel, and Nguyen, 2001)
- CLPRM (b) better than PTRM (a)(based on dic.) whith long queries (Saralegi and Lopez de Lacalle, 2009)
- PTRM (a) independent of retrieval models.

Proposed query translation method

- **Dictionary based** and **parallel corpora free** PTRM:
  - **OOV:** cognate detection on target collection
  - **MWE:** matching and translating by means of MWE lists
  - **Translation selection:** Target collection's co-occurrence based method (Monz and Dorr, 2005)

Outline

1. Introduction
2. Related work
   - Different Strategies
   - CLIR Frameworks based on query translation
3. Proposed query translation method
   - Experimental setup
   - Treatment of OOV words
   - MWE
   - Translation Selection
4. Evaluation
5. Conclusions

Experimental setup

- Topics and Collections:
  - **Development:** CLEF (41-90) topics, LA Times 94 collection, and corresponding HRJ (Human Relevance Judgements)
  - **Test:** CLEF (250-350) topics, LA Times 94 and Glasgow Herald 95 collections, and corresponding HRJ
- Retrieval model: Indri
- Dictionaries:
  - Morris Basque/English dictionary: 77,864 entries and 28,874
  - Euskalterm terminology bank: 72,184 entries and 56,745 unique Basque terms.

Outline

Treatment of OOV words

- Transliteration rules + LCSR:

OOV word	Trans. Rule	Transliteration	Max. LCSR
Txetxenia	tx/ch	chechenia	(chechenia,chechnya)=0.89
korrupzio	-zio/-tion , k/c	corruption	(corruption,corruption)=1

Table: Example of an OOV word resolved using cognate detection

- A total of 64 OOV terms were quantified out and they account for the 15.46% of all query terms
- Most of the OOV words are NEs

Named Entities	Nouns	Adj.	Numbers
82.81%	12.5%	3.13%	1.56%

Table: Distribution of OOV words depending on their POS

Treatment of OOV words

- Cognate based method solves 80% of OOV words
- However, only 7 cases need transliteration and LCSR
- Despite this, 8.96% and 3.52% MAP improvement regarding to baseline (no transliteration and LCSR)
- OOV words tend to be relevant
- We estimated the MAP topline by providing the translations of the OOV words by hand
- **Topline** MAP: translation by hand of all OOV terms
  - 12.38% (short queries), 4.101% (long queries)

Treatment of OOV words

Translation Method	MAP		Improvement Over First %	
	Short	Long	Short	Long
First Translation	0.2703	0.3835		
Topline: First Translation + OOV (by hand)	0.3085	0.3999	12.38	4.101
First Translation + Cognates	0.2969	0.3975	**8.96**	**3.52**

Table: Retrieval performance for OOV words for development topics (41-90 topics)

Outline

MWE

- Treatment: detection on the source query and translation by using a terminology bank
- We identified by hand MWEs on queries:
  - 60 MWEs
  - 51 of them compositional (can be translated word by word)

Basque MWE	Words	Trans. from dic.	Correct candidate
Bigarren Mundu Gerra	Bigarren	second,secondary	second
	Mundu	people, world	world
	Gerra	war	war

Table: Example of word-by-word MWE translation

MWE

- The matching method identifies and translates only 11 MWEs (2 non-compositional)
- Poor coverage but some improvement on MAP terms
  - 5.49 % (short queries), 2.76% (long queries)
  - Most of MWEs compositional $\rightarrow$ translation selection can solve them
- **Topline** MAP: translation by hand of all MWEs
  - 19.81% (short queries), 9.17% (long queries)

MWE

Translation Method	MAP		Improvement Over First %	
	Short	Long	Short	Long
First Translation	0.2703	0.3835		
Topline: First Translation + MWE (by hand)	0.3371	0.4222	19.81	9.17
First Translation + MWE	0.2860	0.3944	**5.49**	**2.76**

Table: Retrieval performance for MWEs for 41-90 topics

Outline

1. Introduction
2. Related work
   - Different Strategies
   - CLIR Frameworks based on query translation
3. Proposed query translation method
   - Experimental setup
   - Treatment of OOV words
   - MWE
   - Translation Selection
4. Evaluation
5. Conclusions

Translation Selection

- Target co-occurrence based selection algorithm:
  - **Idea:** Among all candidates of the source query terms given by the dictionary, select those ones that maximize the global asociation degree between them
- NP-hard maximization problem $\longrightarrow$ Greedy approach (Monz and Dorr, 2005)
  - Initially, all translation candidates are equally likely:

$$w_T^0(t|s_i) = \frac{1}{|tr(s_i)|}$$

  - In the iteration step, each translation candidate is iteratively updated:

$$w_T^n(t|s_i) = w_T^{n-1}(t|s_i) + \sum_{t' \in inlink(t)} \mathbf{w_L}(\mathbf{t}, \mathbf{t'}) * w_T(t'|s_i)$$

Translation Selection

- Measuring Association degree $(\mathbf{w_L}(\mathbf{t}, \mathbf{t'}))$
  - Log-likelihood Ratio (LLR) and co-occurrences between lemmas
  - LLR+nearness factor: Including the distance between source words
  - Log-likelihood Ratio (LLR) and co-occurrences between expanded lemmas

Translation Selection

- AM including distance (formula)

$$w'_L(t,t') = w_L(t,t') * w_F(t,t')$$

$$w'_F(t,t') = \frac{max_{s_i,S_j \in Q} dis(S_i,S_j)}{dis(so(t),so(t'))} * 2^{smw(so(t),so(t'))}$$

  - Strong evidence, more weight (formula):

$$smw(s,s') = \begin{cases} 1 & \text{if } \{s,s'\} \subseteq Z \text{ where } Z \in MWE \\ 0 & \text{else} \end{cases}$$

Translation Selection

- Association between expanded tokens
  - $\mathbf{S_1}$: Source query word 1.
  - $\mathbf{S_2}$: Source query word 2.
  - $\mathbf{C_1}$ and $\mathbf{C_2}$: Senses for source query word 1.
  - $\mathbf{C_3}$: Sense for source query word 2.
  - $\mathbf{t_1}$ and $\mathbf{t_2}$: Trans. candidates for sense C_1.
  - $\mathbf{t_3}$: Trans. candidates for sense C_2.
  - Frequency of the senses:

$$f(C_x) = \sum_{t \in C_x} f(t)$$

  - Frequency between senses:

$$f(C_1 \cap C_3) = f((\cup_{t \in C_1} t) \cap (\cup_{t \in C_3} t))$$

Translation Selection

- **Toplines:** by hand
  - Select the correct translation from candidates of MRD
    - 21.19% (short queries), 10.10% (long queries)
  - If no candidate, take it from english monolingual
    - 32.49% (short queries), 16.50% (long queries)

Translation Selection

Translation Method	MAP		Improvement Over First %	
	Short	Long	Short	Long
First Translation	0.2703	0.3835		
Topline 1: translation Selection by hand	0.3430	0.4266	21.19	10.10
Target co-occurrence based	0.3405	0.4123	20.62	6.99
Topline 2: translation Selection by hand + new translations	0.4004	0.4593	32.49	16.50
Target co-occurrence based + nearness	0.3399	0.4117	20.48	6.85
Target co-occurrence (expanded tokens)	0.3323	0.4163	18.05	7.88

Table: Retrieval perfomance for translation selection for development topics (41-90 topics)

Outline

1. Introduction
2. Related work
   - Different Strategies
   - CLIR Frameworks based on query translation
3. Proposed query translation method
   - Experimental setup
   - Treatment of OOV words
   - MWE
   - Translation Selection
4. Evaluation
5. Conclusions

Evaluation

- Runs:
  - English monolingual (topline)
  - First translation from the dictionary (baseline)
  - OOV: First trans. and cognate detection
  - MWE: MWE translation and First trans.
  - TS: Co-occurrence-based translation selection
  - TS+Nearness: including the nearness factor
  - TS (expanded tokens): Sense co-occurrence
  - TS (expanded tokens)+OOV
  - TS (expanded tokens)+OOV+MWE

Evaluation: Independent Methods

- Runs:
  - English monolingual (topline)
  - First translation from the dictionary (baseline)
  - OOV: First trans. and cognate detection
  - MWE: MWE translation and First trans.
  - TS: Co-occurrence-based translation selection
  - TS+Nearness: including the nearness factor
  - TS (expanded tokens): Sense co-occurrence
  - TS (expanded tokens)+OOV
  - TS (expanded tokens)+OOV+MWE

Evaluation Results: Independent Methods

Run	**MAP**		**% of Monolingual**		**Improvement Over First %**	
	Short	**Long**	**Short**	**Long**	**Short**	**Long**
English monolingual	0.3176	0.3773				
Baseline	0.2195	0.2599	67	69		
OOV	0.2279	**0.2670**	72	71	**7.24**	**2.66**
MWE	0.2237	0.2601	70	69	5.5	0.08

Table: MAP values for test topics (250-350)

Evaluation: Independent Methods

- Runs:
  - English monolingual (topline)
  - First translation from the dictionary (baseline)
  - OOV: First trans. and cognate detection
  - MWE: MWE translation and First trans.
  - TS: Co-occurrence-based translation selection
  - TS+Nearness: including the nearness factor
  - TS (expanded tokens): Sense co-occurrence
  - TS (expanded tokens)+OOV

Evaluation Results: Independent Methods

Run	MAP		% of Monolingual		Improvement Over First %	
	Short	Long	Short	Long	Short	Long
English monolingual	0.3176	0.3773				
Baseline	0.2195	0.2599	67	69		
OOV	0.2279	**0.2670**	72	71	**7.24**	**2.66**
MWE	0.2237	0.2601	70	69	5.5	0.08
TS	0.2315	0.2642	73	70	**8.68**	1.63
TS+Nearness	**0.2318**	0.2627	73	70	**8.8**	1.07
TS (expanded tokens)	0.2362	0.2747	74	73	10.5	5.39

Table: MAP values for test topics (250-350)

Evaluation: Method Combinations

- Topics and collections:
  - **Test:** CLEF (250-350) topics, LA Times 94 and Glasgow Herald 95 collections, and corresponding HRJ
- Runs:
  - English monolingual (topline)
  - First translation from the dictionary (baseline)
  - OOV: First trans. and cognate detection
  - MWE: MWE translation and First trans.
  - TS: Co-occurrence-based translation selection
  - TS+Nearness: including the nearness factor
  - TS (expanded tokens): Sense co-occurrence
  - TS (expanded tokens)+OOV

Evaluation Results: Method Combinations

Run	MAP		% of Monolingual		Improvement Over First %	
	Short	Long	Short	Long	Short	Long
English monolingual	0.3176	0.3773				
Baseline	0.2195	0.2599	67	69		
OOV	0.2279	**0.2670**	72	71	**7.24**	**2.66**
MWE	0.2237	0.2601	70	69	5.5	0.08
TS	0.2315	0.2642	73	70	**8.68**	1.63
TS+Nearness	**0.2318**	0.2627	73	70	**8.8**	1.07
TS (expanded tokens)	0.2362	0.2747	74	73	10.5	5.39
TS (expanded tokens)+OOV	**0.2424**	**0.2805**	76	74	**12.79**	**7.34**

Table: MAP values for test topics (250-350)

Evaluation Results

- Co-occcurrences based method and cognate detection based method improve the baseline significantly
- Expanded token co-occurrences better than token co-occurrences
- MWE treatment poor due to lack of recall

Outline

1. Introduction
2. Related work
   - Different Strategies
   - CLIR Frameworks based on query translation
3. Proposed query translation method
   - Experimental setup
   - Treatment of OOV words
   - MWE
   - Translation Selection
4. Evaluation
5. Conclusions

Conclusions

- Translation selection (including non-compositional MWE) **decreases MAP the most** on a dictionary-based approach
  - Wrong selection (10% short queries, 21% long queries)
  - Wrong selection+No correct translation on MRD (17% queries, 32% queries)
- **OOV terms the least influential** factor (12% queries, 4% queries)
- **Proposed** dictionary-based parallel corpora free **methods offer significant improvement**
  - Co-occurrence based translation selection algorithm
  - Cognate detection method

References I

Darwish, K. and D.W. Oard. 2003. Probabilistic structured query methods. In In Proceedings of the 21st Annual 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 338–344. ACM.

Gao, Jianfeng, Ming Zhou, Jian-Yun Nie, Hongzhao He, and Weijun Chen. 2002. Resolving query ambiguity using a decaying co-occurrence model and syntactic dependence relations. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pages 183–190. ACM.

References II

Hull, D.A. and G. Grefenstette. 1996. Querying across languages: a dictionary-based approach to multilingual information retrieval. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pages 49–57. ACM.

McCarley, J. Scott. 1999. Should we translate the documents or the queries in cross-language information retrieval? In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics 1999, pages 208–214. Association for Computational Linguistics.

References III

Monz, C. and B.J. Dorr. 2005. Iterative translation disambiguation for cross-language information retrieval. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pages 520–527. ACM.

Oard, Douglas W. 1998. A comparative study of query and document translation for cross-language information retrieval. In AMTA '98: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup, pages 472–483, London, UK. Springer-Verlag.

Pirkola, A. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 55–63.

References IV

Saralegi, Xabier and Maddalen Lopez de Lacalle. 2010. Aestimating translation probabilities from the web for structured queries on clir. In In Proceedings of the 32th European Conference on Information Retrieval, pages 586–589. Springer.

Saralegi, Xabier and Maddallen Lopez de Lacalle. 2009. Comparing different approaches to treat translation ambiguity in clir: Structured queries vs. target co-occurrences based selection. In In Proceedings of the 6th international workshop on Text-based Information Retrieval, pages 398 – 404.

Wang, Jianqiang and Douglas W. Oard. 2003. Combining query translation and document translation in cross-language retrieval. In Proceedings of the 4th Workshop of the Cross-Language Evaluation Forum, pages 108–121. Springer Berlin / Heidelberg.

References V

Xu, Jinxi, Ralph Weischedel, and Chanh Nguyen. 2001. Evaluating a probabilistic model for cross-lingual information retrieval. In SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pages 105–110, New York, NY, USA. ACM.

Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR

Xabier Saralegi Maddalen López de Lacalle

R&D
Elhuyar Foundation

7th international conference on Language Resources and Evaluation
LREC 2010, Valletta, Malta
2010/05/20