ResPubliQA - QA@CLEF 2009

Semantic relatedness and cross-lingual passage retrieval

Eneko Agirre[1], Olatz Ansa[1], Xabier Arregi[1], Maddalen Lopez de Lacalle[2], **Arantxa Otegi**[1], Xabier Saralegi[2], Hugo Zaragoza[3]

[1] IXA NLP Group, University of the Basque Country
[2] R&D, Elhuyar Foundation, Basque Country
[3] Yahoo! Research, Barcelona

Introduction

- We participated in...
  - English-English monolingual (EN-EN)
  - Basque-English cross-lingual (EU-EN)
- Our focus:
  - Check IR only for passage retrieval (no question analysis or answer validation)
  - Check Machine Readable Dictionary (MRD) techniques for the EU-EN
  - Check WordNet-based semantic relatedness to expand the passages

English-English (EN-EN)

- No question analysis
- Passage retrieval:

  expansion of passage terms based on related concepts
- No answer validation

English-English (EN-EN)

- No question analysis
- Passage retrieval:

  expansion of passage terms based on related concepts
- No answer validation

Basque-English (EU-EN)

- No question analysis, but
  - Question pre-processing:
    - lemmatize, POS tagging, named entity recognition
  - Translation of query terms to English
- Passage retrieval:

  expansion of passage terms
  based on related concepts
- No answer validation

Basque-English (EU-EN)

- No question analysis, but
  - Question pre-processing:
    - lemmatize, POS tagging, named entity recognition
  - Translation of query terms to English
- Passage retrieval:

  expansion of passage terms
  based on related concepts
- No answer validation

Translation of query terms

- From Basque to English
- No Basque version of document collection
- Strategy:
  - for each keyword take all the translation candidates from two Basque-English MRD
  - for out-of-vocabulary words
    - search for cognates in the target collection
  - ambiguous translations
    - translation selection: co-occurrence optimization (Monz&Dorr)

Passage retrieval

- Split the documents into paragraphs
- Lemmatize and PoS tag passages
- Expand the documents based on semantic relatedness
  - UKB: publicly available graph-based WSD and lexical relatedness engine (Agirre et al. 2009)
  - Given a passage, UKB returns a vector of scores for concepts in WordNet, with most related at top
  - Expand the highest-scoring 100 concepts to all their variants

Passage retrieval

- Index the passages using MG4J
  - one index for the original words and one for the expanded words
  - porter stemmer
- BM25 ranking function
  - we did not tune the k_1 and b parameters
- Return just the 1st passage

Results

submitted runs		#answered correctly	#answered incorrectly	c@1
English-English	run1	211	289	0.42
	run2	240	260	**0.48**
Basque-English	run1	78	422	0.16
	run2	90	409	**0.18**

- run1: not using expansion
- run2: using expansion
  - Semantic relatedness improves results in both tasks, but below baseline

Example of a document expansion

- question (no. 32): *Into which plant may genes be introduced and not raise any doubts about unfavourable consequences for people's health?*

Example of a document expansion

- question (no. 32): *Into which plant may genes be introduced and not raise any doubts about unfavourable consequences for people's health?*

original passage:

Whereas the Commission, having examined each of the objections raised in the light of Directive 90/220/EEC, the information submitted in the dossier and the opinion of the Scientific Committee on Plants, has reached the conclusion that there is no reason to believe that there will be any ***adverse effects*** *on human health or the environment from the introduction into maize of the gene coding for phosphinotricine-acetyl-transferase and the truncated gene coding for beta-lactamase;*

Example of a document expansion

- question (no. 32): *Into which plant may genes be introduced and not raise any doubts about unfavourable consequences for people's health?*

original passage:

Whereas the Commission, having examined each of the objections raised in the light of Directive 90/220/EEC, the information submitted in the dossier and the opinion of the Scientific Committee on Plants, has reached the conclusion that there is no reason to believe that there will be any ***adverse effects*** *on human health or the environment from the introduction into maize of the gene coding for phosphinotricine-acetyl-transferase and the truncated gene coding for beta-lactamase;*

some expanded words:

cistron factor gene coding cryptography ... acetyl acetyl_group acetyl_radical ethanoyl_group ethanoyl_radical beta_lactamase penicillinase common_market ec eec eu europe european_community european_economic_community european_union ... directive directing directional guiding citizens_committee committee environment surround surroundings corn indian_corn maize zea_mays health wellness health adverse contrary homo human human_being man adverse inauspicious untoward lemon lemon_yellow ... unfavorable ***unfavourable*** *... set_up expostulation objection remonstrance remonstration dissent protest believe light lightly belief feeling impression notion opinion ... reason reason_out argue jurisprudence law* ***consequence*** *effect event issue outcome result ...*

Analysis

- Performance drops in the Basque-English task
  - 38% of monolingual, when same technique achieves 74% in other settings
- Basque has no reference document collection or reference terminology for this domain
  - “Official Journal of the Community”
- Many query/answer pairs in the other languages were literal
- Unfortunately, no other cross-lingual participant

Example

- EU: *Nola izendatuko ditu Kontseiluak epaileak?*

Example

- EU: *Nola izendatuko ditu Kontseiluak epaileak?*
- EN: *How **will judges be appointed by the Council**?*

Example

- EU: *Nola izendatuko ditu Kontseiluak epaileak?*
- EN: *How **will judges be appointed by the Council**?*

- *<answer_english_string e_doc_id="jrc32005D0150-en" e_p_id="32">The **judges will be appointed by the Council** acting unanimously, after consulting the committee of seven persons chosen from among former members of the Court of Justice and the Court of First Instance and lawyers of recognised competence. The committee will give its opinion on the candidates' suitability to perform the duties of judge at the Civil Service Tribunal ...</answer_english_string>*

Example

- EU: *Nola izendatuko ditu Kontseiluak epaileak?*
- EN: *How **will judges be appointed by the Council**?*
- EU keywords: *izendatu kontseilu epaile*
- Translation to EN: ***designate** council judge*
- *<answer_english_string e_doc_id="jrc32005D0150-en" e_p_id="32">The **judges will be appointed by the Council** acting unanimously, after consulting the committee of seven persons chosen from among former members of the Court of Justice and the Court of First Instance and lawyers of recognised competence. The committee will give its opinion on the candidates' suitability to perform the duties of judge at the Civil Service Tribunal ...</answer_english_string>*

Analysis

- Performance drops in the Basque-English task
  - 38% of monolingual, when same technique achieves XX in other settings
- Basque has no reference document collection or reference terminology for this domain
  - “official journal of the European Commission”
- Many query/answer pairs in the other languages were literal
- Unfortunately, no other cross-lingual participant

Conclusions and future work

- Good results can be achieved without question analysis and answer validation
- Results improve applying semantic relatedness

- Optimize parameters to beat baseline
- Gather comparable corpora to improve cross-lingual results (Talvensaari, 2008)