TNO at CLEF-2001

Comparing Translation Resources

Wessel Kraaij

kraaij@tpd.tno.nl

TNO-TPD

P.O. Box 155, 2600 AD Delft

The Netherlands

Overview

- Research questions

Overview

- Research questions
- Basic (CL)IR approach

Overview

- Research questions
- Basic (CL)IR approach
- Translation resources

Overview

- Research questions
- Basic (CL)IR approach
- Translation resources
- Experimental Results

Overview

- Research questions
- Basic (CL)IR approach
- Translation resources
- Experimental Results
- Conclusions

CLEF 2001 Research questions

- Evaluate individual components of multilingual IR system
- Compare translation resources
- How can we integrate translation into the IR model?
- How can we normalize scores?
- Proper noun recognition

Basic CLIR Approach

- Query translation using dictionaries or MT
- Stop and Lemmatise queries and documents (XRCE tools)
- Focus on integrating translation into retrieval model
- Normalising scores for merging

Retrieval Model

- Language Models for IR gain support (LMIR workshop, CMU, may 2001)
- Ponte(1998), Hiemstra(1998), Miller et al. (1998)
- Advantages:
  - Cleaner model than Okapi BM25 & Lnu.Ltu
  - Effectiveness is competitive
  - Easy to extend for different tasks: CLIR, filtering, summarisation.

IR based on language models

$$P(Q|D_k) = P(T_1, T_2, \cdots, T_n|D_k) = \prod_{i=1}^{n} P(T_i|D_k)$$

basic model: terms are generated independently

- Query: a sequence of terms
- Retrieval: rank documents on the *query likelihood* given each document.

IR based on language models

$$P(Q|D_k) = \prod_{i=1}^{n} [\lambda_i P(T_i|D_k) + (1 - \lambda_i) P(T_i)]$$

smooth by interpolation

- Query: a sequence of terms
- Retrieval: rank documents on the *query likelihood* given each document.

IR based on language models

$$\log P(Q|D_k) = \sum_{i=1}^{n} \log[\lambda_i P(T_i|D_k) + (1-\lambda_i)P(T_i)]$$

Take logs: products become summations

- Query: a sequence of terms
- Retrieval: rank documents on the *query likelihood* given each document.

IR based on language models

$$LLR(Q|D_k) = \frac{P(Q|D_k)}{P(Q|D_k)} = \sum_{i=1}^{n} \log[\frac{\lambda_i P(T_i|D_k) + (1-\lambda_i)P(T_i)}{P(T_i)}]$$

Normalise scores across queries/collection: query content

- Query: a sequence of terms
- Retrieval: rank documents on the *query likelihood* given each document.

IR based on language models

$$LLR'(Q|D_k) = 1/n \sum_{i=1}^{n} \log[\frac{\lambda_i P(T_i|D_k) + (1-\lambda_i)P(T_i)}{P(T_i)}]$$

Normalise scores across queries: query length

- Query: a sequence of terms
- Retrieval: rank documents on the *query likelihood* given each document.

IR based on language models

$$P(S_1, S_2, \cdots, S_n) = \prod_{i=1}^{n} \sum_{j=1}^{m} P(S_i | T_i = t^{(j)}) P(T_i = t^{(j)} | D_k)$$

Integrate Translation into the basic (unsmoothed) model

- Query: a sequence of terms
- Retrieval: rank documents on the *query likelihood* given each document.

Translation Resources (1) VLIS

- Multilingual lexical database (EN/NL/FR/ES/**IT**/DE)
- Contains translations NL$\rightarrow$EN, synonyms etc.
- All translation pairs are simulated via Dutch as interlingua
- Simple word by word lookup, i.e. no phrase lookup!
- Translation probability $P(w_s|w_t)$ is estimated on pseudo frequency information
- Prune “labeled” translations (pejoratives)
- No contextual sensitivity

Translation Resources (2) Systran (MT)

- Used Babelfish web service (babelfish.pl)
- Currently 19 translation pairs involving 8 different languages
- Full sentence translation.
- “Expression Dictionary”

Translation Resources (3) Parallel Web

- 4 corpora mined by RALI & TNO TPD
- IBM Model 1 trained by RALI for each language pair
- Word by word translation
- Evaluated different pruning techniques
- Empirical validation of the IR model extension

Results: CLEF 2000 topics

language pair	m.a.p.	% of baseline	method
EN-EN	0.4164	100	mono
FR-EN	0.3995	95	web corpus
FR-EN	**0.4007**	95	Babelfish
FR-EN	0.2971	71	VLIS
FR-FR	0.4529	100	mono
EN-FR	**0.3680**	82	web corpus
EN-FR	0.3321	73	Babelfish
EN-FR	0.2773	62	VLIS
EN-EN	0.4164	100	mono
IT-EN	0.3309	79	web corpus
IT-EN	**0.3595**	86	Babelfish
IT-EN	0.3119	75	VLIS
IT-IT	0.4808	100	mono
EN-IT	**0.3771**	79	web corpus
EN-IT	0.3564	75	Babelfish
EN-IT	0.3266	69	VLIS

Results: CLEF 2001 topics

language pair	m.a.p.	% of baseline	method
EN-EN	0.5144	100	mono
FR-EN	0.4637	90	web corpus
FR-EN	**0.4735**	92	Babelfish
FR-EN	0.3711	73	VLIS
FR-FR	0.4877	100	mono
EN-FR	0.3642	76	web corpus
EN-FR	0.4039	82	Babelfish
EN-FR	**0.4051**	83	VLIS
EN-EN	0.5144	100	mono
IT-EN	0.3672	71	web corpus
IT-EN	0.3702	72	Babelfish
IT-EN	**0.3780**	73	VLIS
IT-IT	0.4411	100	mono
EN-IT	0.3137	70	web corpus
EN-IT	0.2824	64	Babelfish
EN-IT	**0.3549**	80	VLIS

FR-EN: Per topic analysis

- Proper names: Tchétchénie, l'IRA, La Lettonie
- phrases: tremblement de terre, agent immobilier
- contextual disambiguation: enchères, faim
- related terms: mort

Babelfish across language pairs

FR-EN	92
EN-FR	86
EN-DE	72
EN-IT	64
EN-ES	80

VLIS across language pairs

FR-EN	73
EN-FR	83
EN-DE	81
EN-IT	80
EN-ES	77
NL-EN	79
NL-FR	86
NL-DE	86
NL-IT	76
NL-ES	66

Integrating a probabilistic dictionary

- Pruning is important
- Corpora size matters
- more translations » best translation
- forward probability » equal probability» reverse probability
- CLIR model might have to be refined

Combination runs

method	language pair	m.a.p.	% of baseline
mono	EN-EN	0.5144	100
web corpus	EN-FR	0.4637	90
Babelfish	EN-FR	0.4735	92
VLIS	EN-FR	0.3711	73
corpus&Babelfish	EN-FR	0.4895	96
corpus&VLIS	EN-FR	0.4672	92
VLIS&Babelfish	EN-FR	0.4783	94
VLIS&Babelfish& corpus	EN-FR	0.5032	98

Multilingual Results

run tag	language pair	m. a. p.	method
tnoex3	EN-X	0.2634	VLIS
tnoex4	EN-X	0.2413	Babelfish
tnonx3	NL-X	0.2513	VLIS

Conclusions

- All three resources can achieve good performance

Conclusions

- All three resources can achieve good performance
- “Simple” web resource performs remarkably well

Conclusions

- All three resources can achieve good performance
- “Simple” web resource performs remarkably well
- Lexical coverage main determinant m.a.p.

Conclusions

- All three resources can achieve good performance
- “Simple” web resource performs remarkably well
- Lexical coverage main determinant m.a.p.
- Problems: phrases, proper nouns

Conclusions

- All three resources can achieve good performance
- “Simple” web resource performs remarkably well
- Lexical coverage main determinant m.a.p.
- Problems: phrases, proper nouns
- Topic collection too small for good comparison

Discussion

How can we evaluate translation methods in an experiment where lexical coverage is a controlled variable?