Boosting Cross-Language Retrieval by Learning Bilingual Phrase Associations from Relevance Rankings

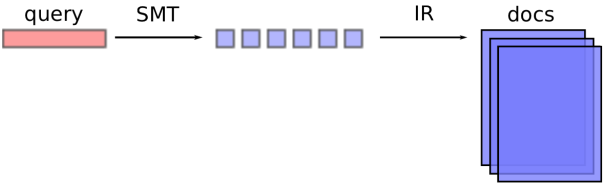
Artem Sokolov, Laura Jehl, Felix Hieber, Stefan Riezler

UNIVERSITÄT HEIDELBERG
ZUKUNFT
SEIT 1386

Cross-Lingual Information Retrieval: State-of-the-art

Direct translation

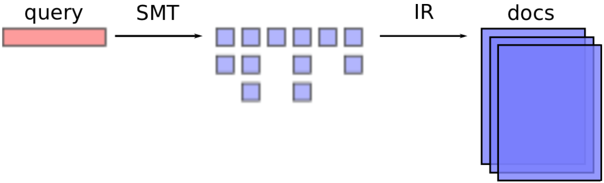
- translate query with SMT system
- monolingual retrieval with 1-best translation
- ✔ easy to deploy
- ✔ useful provided lots of in-domain data



Cross-Lingual Information Retrieval: State-of-the-art

Probabilistic structured queries

- query representation that includes translation alternatives
- estimate expected tf/idf weights & retrieve monolingually
- ✔ uses "good" n-best translations
- ✔ implicit query expansion by considering translation alternatives



Drawbacks of standard approaches

- crucial dependence on SMT quality
- SMT tuned for translation quality
- no learning for retrieval

This paper

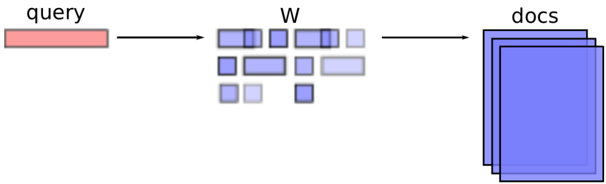
- ✔ learns n-gram "phrase-table" relevant for the task
- ✔ optimizes final retrieval objective
- ✔ independent of any SMT system
- ✔ standalone: as good as a large domain-tuned SMT system or better
- ✔ combined with SMT baselines: +7 MAP & +15 PRES points.

Drawbacks of standard approaches

- crucial dependence on SMT quality
- SMT tuned for translation quality
- no learning for retrieval

This paper

- ✔ learns n-gram "phrase-table" relevant for the task
- ✔ optimizes final retrieval objective
- ✔ independent of any SMT system
- ✔ standalone: as good as a large domain-tuned SMT system or better
- ✔ combined with SMT baselines: +7 MAP & +15 PRES points.



Section 1

Baseline Approaches

Direct translation

(1) SMT model for query translation

- state-of-the-art SCFG decoder (`cdec`) [Dyer 10]
- word alignments from parallel data (`mgiza++`)
- in-domain language model (`kenlm`) [Heafield 11]
- parameter tuning with MERT [Och 03]

(2) Retrieval

- Okapi BM25 ranking

Probabilistic structured queries

(1) Query projection

- **calculate expected tf/idf weights** with word translation probabilities: [Darwish 03, Ture 12]

$$tf(f,E) = \sum_{e \in E} tf(e,E)p(e|f) \qquad df(f) = \sum_{e \in E} df(e)p(e|f)$$

- **estimate** $p(e|f)$**'s from** [Ture 12]
    - lexical translation table
      and/or from
    - word alignments in derivations of the SMT n-best list

(2) Retrieval

- Okapi BM25 ranking

Section 2

Learning Phrase-Tables from Ranking Data

Ranking Approach: Model

- query $\mathbf{q}$, document $\mathbf{d}$ (bag-of-words)

Scoring function

$$f(\mathbf{q},\mathbf{d}) = \mathbf{q}^\top W \mathbf{d} = \sum_{i=1}^{Q} \sum_{j=1}^{D} q_i W_{ij} d_j.$$

Linear model

Assign a weight to every pair of query and document terms:

$$f(\mathbf{q},\mathbf{d}) = \sum_{ij} W_{ij} (\mathbf{q}\mathbf{d}^\top)_{ij} = \mathbf{w}^\top \phi(\mathbf{q},\mathbf{d})$$

Training

Training data

$$\{(\mathbf{q}, \mathbf{d}^+, \mathbf{d}^-)\}$$

where $\mathbf{d}^+$ is a relevant document and $\mathbf{d}^-$ an irrelevant for query $\mathbf{q}$

Task

Find $W \in \mathbb{R}^{Q \times D}$ such that $f(\mathbf{q}, \mathbf{d}^+) > f(\mathbf{q}, \mathbf{d}^-)$ for all training tuples

How to learn big W?

- low-rank decomposition of W [Bai 10]
- force feature selection by ℓ_1-regularization [Chen 10]
- **start from empty W & add features progressively**

Training

Training data

$$\{(\mathbf{q}, \mathbf{d}^+, \mathbf{d}^-)\}$$

where $\mathbf{d}^+$ is a relevant document and $\mathbf{d}^-$ an irrelevant for query $\mathbf{q}$

Task

Find $W \in \mathbb{R}^{Q \times D}$ such that $f(\mathbf{q}, \mathbf{d}^+) > f(\mathbf{q}, \mathbf{d}^-)$ for all training tuples

How to learn big W?

- low-rank decomposition of W [Bai 10]
- force feature selection by ℓ_1-regularization [Chen 10]
- **start from empty W & add features progressively**

Boosting

Exp loss function

$$\mathcal{L}_{exp} = \sum_{(\mathbf{q},\mathbf{d}^+,\mathbf{d}^-)} D(\mathbf{q}, \mathbf{d}^+, \mathbf{d}^-) e^{f^T(\mathbf{q},\mathbf{d}^-) - f^T(\mathbf{q},\mathbf{d}^+)}$$

$D(\mathbf{q}, \mathbf{d}^+, \mathbf{d}^-)$ – importance weighting from relevance levels

Iterative building of scoring function

$$f^T(\mathbf{q}, \mathbf{d}) = \sum_t^T w_{ij}^t q_i^t d_j^t$$

- on step t selects new pair i, j
- D_{t+1} reweighted to concentrate on previously misclassified pairs

An Efficient Implementation of Boosting

Parallelization & Bagging

- each node receives a sample s from training tuples
- when done models are averaged: $f(\mathbf{q}, \mathbf{d}) = \frac{1}{S} \sum_t \sum_s w_t^s h_t^s(\mathbf{q}, \mathbf{d})$

Speed & Memory Tricks

- on-the-fly feature construction (avoids inv. index) [Grangier 08, Goel 08]
- only update gradients for features that cooccur with previously selected one [Collins 05]
- random feature hashing into 2^{30}-sized pool (keep W in RAM) [Shi 09]

An Efficient Implementation of Boosting

Parallelization & Bagging

- each node receives a sample s from training tuples
- when done models are averaged: $f(\mathbf{q}, \mathbf{d}) = \frac{1}{S} \sum_t \sum_s w_t^s h_t^s(\mathbf{q}, \mathbf{d})$

Speed & Memory Tricks

- on-the-fly feature construction (avoids inv. index) [Grangier 08, Goel 08]
- only update gradients for features that cooccur with previously selected one [Collins 05]
- random feature hashing into 2^{30}-sized pool (keep W in RAM) [Shi 09]

Example Learned Phrase-Table

t	h_t **(uni- & bi-grams)**	w_t
1	層$_{\text{layer}}$ - layer	1.29
2	データ$_{\text{data}}$ - data	1.13
3	回路$_{\text{circuit}}$ - circuit	1.13
77	導$_{\text{guide}}$,電$_{\text{power}}$ - conductive	1.25
81	解決$_{\text{resolution}}$ - image	-0.25
99	変速$_{\text{speed}}$ - transmission	1.68
100	液晶$_{\text{LCD}}$ - liquid,crystal	1.73
123	力$_{\text{power}}$ - force	0.91
124	圧縮$_{\text{compression}}$,機$_{\text{machine}}$ - compressor	2.83
132	ケーブル$_{\text{cable}}$ - cable	1.81
133	超$_{\text{hyper}}$,音波$_{\text{sound wave}}$ - ultrasonic	3.34
169	粒子$_{\text{particle}}$ - particles	1.57
170	算出$_{\text{calculation}}$ - for,each	1.14
184	ロータ$_{\text{rotor}}$ - rotor	2.01
185	検出$_{\text{detection}}$,器$_{\text{vessel}}$ - detector	1.43

Section 3

Experiments

Parallel Translation Data (JP→EN)

Training

NTCIR-7 PatentMT workshop data (1.8M sentences)

Parameter tuning

parameter tuning: NTCIR-8 test collection (2K sentences)

Ranking Data

Automatic extraction of relevance judgements [Graf 08]

- cross-language citation graph from MAREC corpus to extract patents in citation or family relation
- 3 relevance levels:
  - **3** family patents (same invention granted elsewhere)
  - **2** cited by examiners
  - **1** cited by applicants

- extracted abstracts from MAREC and NTCIR-10

	queries	relevant docs
train	100k	1.5M
dev	2k	26k
test	2k	25k

http://www.cl.uni-heidelberg.de/statnlpgroup/boostclir

Performance of standalone systems

- MAP - Mean Average Precision
- PRES - Patent Retrieval Evaluation Score (recall-oriented) [Magdy 11]
- both $\in [0,1]$; higher is better

	test MAP	test PRES
DT[1]	0.2555	0.5681
PSQ lexical table[2]	0.2444	0.5498
PSQ n-best table[3]	**0.2659**	0.5851
Boost-unigram	[1,2,3]0.1982	[1,2]**0.6122**
Boost-bigram	[3]**0.2474**	[1,2,3]**0.7196**

- small boosting models: ∼100K (1-gram) & ∼170K (2-gram)
- lexical table: ∼600K entries

Rank Aggregation

Intuition

- SMT helpful for cohesive, general passages
- Boosting provides task-specific info complementary to SMT:
  - ✔ rewards phrase pairs that aid retrieval and
  - ✔ penalizes pairs that are detrimental to the task

Best of both worlds

- aggregate systems with orthogonal information sources
- consensus voting (Borda Count) + interpolation:

$$f_{agg}(\mathbf{q}, \mathbf{d}) = \kappa \frac{f_1(\mathbf{q}, \mathbf{d})}{\sum_{\mathbf{d}} f_1(\mathbf{q}, \mathbf{d})} + (1 - \kappa) \frac{f_2(\mathbf{q}, \mathbf{d})}{\sum_{\mathbf{d}} f_2(\mathbf{q}, \mathbf{d})}$$

Rank Aggregation

Intuition

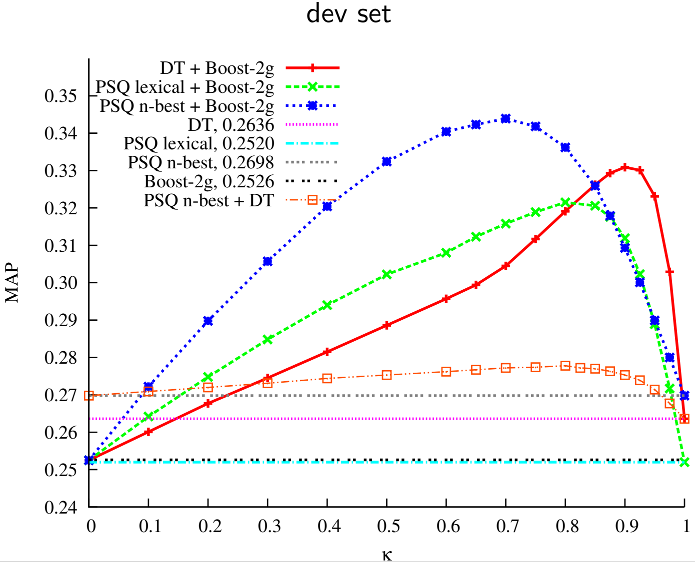
- SMT helpful for cohesive, general passages
- Boosting provides task-specific info complementary to SMT:
  - ✔ rewards phrase pairs that aid retrieval and
  - ✔ penalizes pairs that are detrimental to the task

Best of both worlds

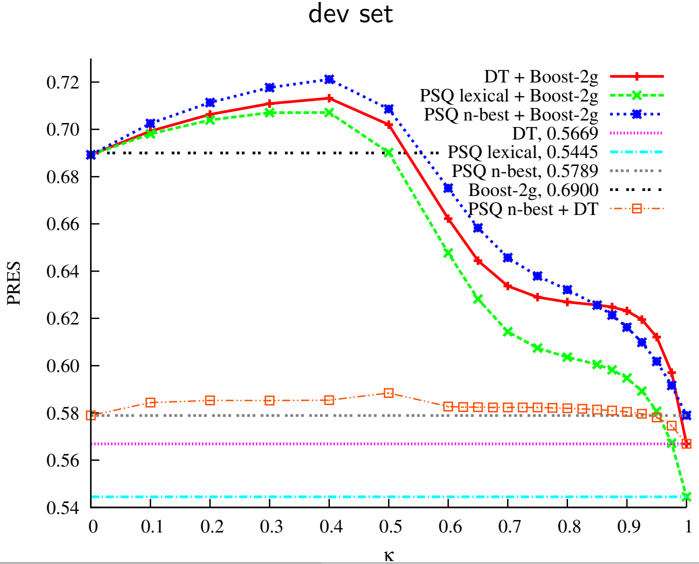
- aggregate systems with orthogonal information sources
- consensus voting (Borda Count) + interpolation:

$$f_{agg}(\mathbf{q}, \mathbf{d}) = \kappa \frac{f_1(\mathbf{q}, \mathbf{d})}{\sum_{\mathbf{d}} f_1(\mathbf{q}, \mathbf{d})} + (1 - \kappa) \frac{f_2(\mathbf{q}, \mathbf{d})}{\sum_{\mathbf{d}} f_2(\mathbf{q}, \mathbf{d})}$$

Performance of aggregated systems: MAP



Performance of aggregated systems: PRES



Performance aggregated systems: overall

method	test MAP	test PRES
DT + PSQ n-best	*0.2726	*0.5942
DT + Boost-1g	*0.2728	*0.6225
DT + Boost-2g	***0.3300**	***0.7279**
PSQ lexical + Boost-1g	*0.2653	*0.6131
PSQ lexical + Boost-2g	***0.3187**	***0.7240**
PSQ n-best + Boost-1g	*0.2850	*0.6402
PSQ n-best + Boost-2g	***0.3416**	***0.7376**

- aggregating two SMT-based systems does not help!
- aggregating orthogonal systems gives up to **+7 MAP/+15 PRES**

Take-away message

- ✔ encode task-relevant information into “phrase-table”
- ✔ orthogonal & complementary information to standard CLIR
- ✔ aggregation with standard SMT gives a huge boost in performance

Data available at

http://www.cl.uni-heidelberg.de/statnlpgroup/boostclir

Take-away message

- ✔ encode task-relevant information into "phrase-table"
- ✔ orthogonal & complementary information to standard CLIR
- ✔ aggregation with standard SMT gives a huge boost in performance

Data available at

`http://www.cl.uni-heidelberg.de/statnlpgroup/boostclir`

Thank you

Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle & Kilian Weinberger.
Learning to Rank with (a Lot of) Word Features.
Information Retrieval Journal, vol. 13, no. 3, pages 291–314, 2010.

Xi Chen, Bing Bai, Yanjun Qi, Qihang Ling & Jaime Carbonell.
Learning Preferences with Millions of Parameters by Enforcing Sparsity.
In Proceedings of the IEEE International Conference on Data Mining (ICDM'10), Sydney, Australia, 2010.

Michael Collins & Terry Koo.
Discriminative Reranking for Natural Language Parsing.
Computational Linguistics, vol. 31, no. 1, pages 25–69, 2005.

Kareem Darwish & Douglas W. Oard.
Probabilistic Structured Query Methods.
In Proceedings. of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03), Toronto, Canada, 2003.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman & Philip Resnik.
cdec: A Decoder, Alignment, and Learning Framework for Finite-State and Context-Free Translation Models.
In Proceedings of the ACL 2010 System Demonstrations, Uppsala, Sweden, 2010.

Sharad Goel, John Langford & Alexander L. Strehl.
Predictive Indexing for Fast Search.
In Advances in Neural Information Processing Systems, Vancouver, Canada, 2008.

Erik Graf & Leif Azzopardi.
A methodology for building a patent test collection for prior art search.
In Proceedings of the 2nd International Workshop on Evaluating Information Access (EVIA), Tokyo, Japan, 2008.

David Grangier & Samy Bengio.
A discriminative kernel-based approach to rank images from text queries.
IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 30, no. 8, pages 1371–1384, 2008.

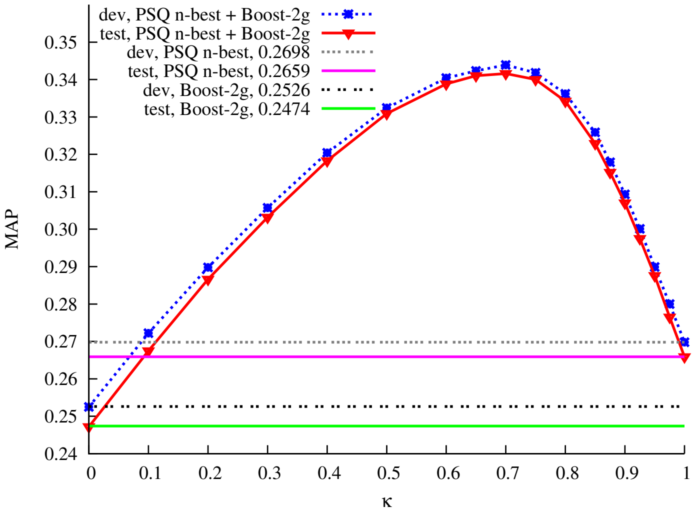
Kenneth Heafield.
KenLM: Faster and Smaller Language Model Queries.
In Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation (WMT'11), Edinburgh, UK, 2011.

Walid Magdy & Gareth J. F. Jones.
An Efficient Method for Using Machine Translation Technologies in Cross-Language Patent Search.
In Proceedings of the 20th ACM Conference on Informationand Knowledge Management (CIKM'11), Glasgow, Scotland, UK, 2011.

Franz Josef Och.
Minimum error rate training in statistical machine translation.
In Proceedings of the 41st Meeting on Association for Computational Linguistics (ACL'03), Sapporo, Japan, 2003.

Qinfeng Shi, James Petterson, Gideon Dror, John Langford, Alexander J. Smola, Alexander L. Strehl & Vishy Vishwanathan.
Hash Kernels.
In Proceedings of the 12th Int. Conference on Artificial Intelligence and Statistics (AISTATS'09), Irvine, CA, 2009.

Ferhan Ture, Jimmy Lin & Douglas W. Oard.
Looking Inside the Box: Context-Sensitive Translation for Cross-Language Information Retrieval.
In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012), Portland, OR, 2012.



Verification that gains transfer to the test data.