# Bag-of-Words Forced Decoding for Cross-Lingual Information Retrieval

**Felix Hieber**
Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
hieber@cl.uni-heidelberg.de

**Stefan Riezler**
Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
riezler@cl.uni-heidelberg.de

## Abstract

Current approaches to cross-lingual information retrieval (CLIR) rely on standard retrieval models into which query translations by statistical machine translation (SMT) are integrated at varying degree. In this paper, we present an attempt to turn this situation on its head: Instead of the retrieval aspect, we emphasize the translation component in CLIR. We perform search by using an SMT decoder in forced decoding mode to produce a bag-of-words representation of the target documents to be ranked. The SMT model is extended by retrieval-specific features that are optimized jointly with standard translation features for a ranking objective. We find significant gains over the state-of-the-art in a large-scale evaluation on cross-lingual search in the domains patents and Wikipedia.

## 1 Introduction

Approaches to CLIR have been plentiful and diverse. While simple word translation probabilities are easily integrated into term-based retrieval models (Berger and Lafferty, 1999; Xu et al., 2001), state-of-the-art SMT systems (Koehn, 2010; Chiang, 2007) are complex statistical models on their own. The use of established translation models for context-aware translation of query strings, effectively reducing the problem of CLIR to a pipeline of translation and monolingual retrieval, has been shown to work well in the past (Chin et al., 2008). Only recently, approaches have been presented to include (weighted) translation alternatives into the query structure to allow a more generalized term matching (Ture et al., 2012a; Ture et al., 2012b). However, this integration of SMT remains agnostic about its use for CLIR and is instead optimized to match fluent, human reference translations. In contrast, retrieval systems often use bag-of-word representations, stopword filtering, and stemming techniques during document scoring, and queries are rarely fluent, grammatical natural language queries (Downey et al., 2008). Thus, most of a translation's structural information is lost during retrieval, and lexical choices may not be optimal for the retrieval task. Furthermore, the nature of modeling translation and retrieval separately requires that a single query translation is selected, which is usually done by choosing the most probable SMT output.

Attempts to inform the SMT system about its use for retrieval by optimizing its parameters towards a retrieval objective have been presented in the form or re-ranking (Nikoulina et al., 2012) or ranking (Sokolov et al., 2014). In this paper, we take this idea a step further and directly integrate the task of scoring documents with respect to the query into the process of *translation decoding*. We make the full expressiveness of the translation search space available to the retrieval model, without enumerating all possible translation alternatives. This is done by augmenting the linear model of the SMT system with features that relate partial translation hypotheses to documents in the retrieval collection. These retrieval-specific features decompose over partial translation hypotheses and thus allow efficient decoding using standard dynamic programming techniques. Furthermore, we apply learning-to-rank to jointly optimize translation and retrieval for the objective of retrieving relevant documents, and use decoding over the weighted translation hypergraph directly to perform cross-lingual search. Since high weights on retrieval features for words in the bag-of-words (BOW) representation of documents *force* the decoder to prefer relevant documents with high probability, by a slight abuse of terminology, we call our approach *BOW Forced Decoding*.

One of the key features of our approach is the use of context-sensitive information such as the language model and reordering information. We show that the use of such a translation-benign search space is crucial to outperform state-of-the-art CLIR approaches. Our experimental evaluation of retrieval performance is done on Wikipedia cross-lingual article retrieval (Bai et al., 2010; Schamoni et al., 2014) and patent prior art search (Fujii et al., 2009; Guo and Gomes, 2009; Sokolov et al., 2013; Schamoni et al., 2014). On both datasets, we show substantial improvements over the CLIR baselines of direct translation (Chin et al., 2008) or Probabilistic Structured Queries (Ture et al., 2012b), with and without further parameter tuning using learning-to-rank techniques and extended feature sets. From our results we conclude, that, in spite of algorithmic complexity, it is central to model translation and retrieval jointly to create more powerful CLIR models.

## 2 Related Work

The framework of translation-model based retrieval has been introduced by Berger and Lafferty (1999). An extension to the cross-lingual case using context-free lexical translation tables has been given by Xu et al. (2001). While the industry standard to CLIR is a pipeline of SMT-based query translation feeding into monolingual retrieval (Chin et al., 2008), recent approaches include (weighted) SMT translation alternatives into the query structure to allow a more generalized term matching (Ture et al., 2012a; Ture et al., 2012b). Less work has been devoted to optimizing SMT towards a retrieval objective, for example in a re-ranking framework (Nikoulina et al., 2012) or by integrating a decomposable proxy for retrieval quality of query translations into discriminative ranking (Sokolov et al., 2014).

The idea of forced decoding has been employed recently to select better perceptron updates from the full SMT search space for discriminative parameter tuning of SMT systems (Yu et al., 2013; Zhao et al., 2014).

Most similar to our approach is the recent work of Dong et al. (2014) who use the `Moses` translation option lattices for translation retrieval, i.e., for mining comparable data. Their query lattices given by the translation options encode exponentially many queries and are used to retrieve the most probable translation candidate from a set of candidates. The approach is evaluated in the context of a parallel corpus mining system. We present a model that not only uses the full search space, including the language model and reordering information, but also evaluate the model specifically for the task of retrieval, rather than mate-finding only. We show that a forced decoding model using bag-of-word representations for documents and retrieval features that are decomposable over query terms significantly outperforms state-of-the-art CLIR baselines such as direct translation (Chin et al., 2008) or Probabilistic Structured Queries obtained from $n$-best list query translations (Darwish and Oard, 2003; Ture et al., 2012b). Additionally we find that the use of context-sensitive translation information such as language models or reordering information, greatly improves retrieval quality in these types of models. We furthermore show how to directly optimize the retrieval objective using large-scale retrieval data sets with automatically induced relevance judgments.

## 3 A Bag-of-Words Forced Decoding Model

**Model Definition.** SMT systems use a Viterbi approximation to find the output hypothesis $q_e^*$

$$q_e^* = \arg\max_{q_e} \max_{h \in \mathcal{E}_{q_f}} P(h, q_e | q_f). \qquad (1)$$

over the search space of hypotheses or derivations $h \in \mathcal{E}_{q_f}$ for a given input $q_f$. The probability of a translation output $q_e$ under derivation $h$ given $q_f$ is usually modeled in a log-linear model

$$P(h, q_e | q_f; \mathbf{w}_{smt}) = \frac{e^{F_{smt}(h, q_e, q_f)}}{\sum_{q_e, h} e^{F_{smt}(h, q_e, q_f)}},$$

where $F(h, q_e, q_f)$ is a learned linear combination of input-output features, that is, the dot product between parameter column vector $\mathbf{w}_{smt}$ and feature column vector given by feature map $\mathbf{\Phi}_{smt}$,

$$F_{smt}(h, q_e, q_f) = \mathbf{w}_{smt}^T \mathbf{\Phi}_{smt}(h, q_e, q_f). \quad (2)$$

In CLIR, we seek to choose a derivation that is *both* an accurate translation of the input according to the translation model, and a well-formed discriminative query that matches relevant documents with high probability. We combine both objectives by directly modeling the probability of a document $d_e$ in target language $e$ given a query $q_f$ in source language $f$, factorized as follows:

$$P(d_e|q_f) = \sum_{h \in \mathcal{E}_{q_f}} \underbrace{P(h|q_f)}_{\text{translation}} \times \underbrace{P(d_e|h, q_f)}_{\text{retrieval}}.$$

Applying the same Viterbi approximation during inference as in (1), we choose the retrieval score of $d_e$ to be the score of the highest scoring hypothesis $h$,

$$score(q_f, d_e) = \max_{h \in \mathcal{E}_{q_f}} P(h|q_f) \times P(d_e|h, q_f), \; (3)$$

where the product between both models can be interpreted as a conjunctive operation similar to a product of experts (Hinton, 2002): A high score is achieved if both experts, namely translation and retrieval models, assign high scores to a hypothesis. That is, the model attempts to produce a well-formed translation, but at the same time chooses lexical items present in the bag-of-words representation of the document. Similarly, we can interpret the inclusion of the retrieval component as a constraint to *force* the decoder to retrieve $d_e$ with high probability. By a slight abuse of terminology, we will henceforth call our approach Bag-of-Words Forced Decoding (BOW-FD).[1]

The translation term $P(h|q_f)$ is modeled as in (2) for standard hierarchical phrase-based SMT (Chiang, 2007) and left unchanged in our joint model. The retrieval term $P(d_e|h, q_f)$ is modeled in a similar form

$$F_{ir}(h, d_e) = \mathbf{w}_{ir}^T \mathbf{\Phi}_{ir}(h, d_e),$$

where IR features do not depend on $q_f$ (thus allowing us to drop this term) and decompose over derivation terms. This allows a bag-of-word vector representation of documents, and retrieval features are local to single edges in the search space for efficient Viterbi inference. The joint scoring model is defined as follows:

$$score(q_f, d_e; \mathbf{w}) = \max_{h \in \mathcal{E}_{q_f}} e^{F_{smt}(h, q_e, q_f) + F_{ir}(h, d_e)},$$

where the weight vector is defined by the vector concatenation $\mathbf{w} = \mathbf{w}_{smt} \| \mathbf{w}_{ir}$, and $q_e$ refers to the yield that is determined uniquely by derivation $h$.

Following the interpretation of our joint model as forced or constrained decoding, we can view pipeline approaches such as the direct translation baseline as instances of *unconstrained* decoding. That is, the SMT decoder yields a single translation output for every document and the assignment of document scores is deferred to a (monolingual) retrieval model given this single output structure. Other CLIR approaches such as probabilistic structured queries (Darwish and Oard, 2003; Ture et al., 2012b) try to mitigate this early disambiguation by keeping enumerated translation alternatives at retrieval time. However, they either use context-free word-based translation tables or select only terms from a small $n$-best fraction of the full search space.

**Dynamic Programming on Hypergraphs.** Decoding in a hierarchical phrase-based SMT (Chiang, 2007) is usually understood as a two-step process: Initially, an input sentence is parsed using a Weighted Synchronous Context-Free Grammar (WSCFG) in a bottom-up manner to construct an initial hypergraph $\mathcal{H}$ that compactly encodes the full search space ("translation forest") (Gallo et al., 1993; Klein and Manning, 2001; Huang and Chiang, 2005; Dyer et al., 2010). An ordered, directed hypergraph $\mathcal{H}$ is a tuple $\langle V, E, g, \mathcal{W} \rangle$, consisting of a finite set of nodes $V$, a finite set of hyperedges $E$, and weight function $\mathcal{W} : E \mapsto \mathbb{R}$ assigning real-valued weights to $e \in E$. Language models are typically added in a second rescoring phase that is carried out by approximate solutions, such as cube-pruning (Chiang, 2007; Huang and Chiang, 2007), limiting the number of derivations created at each node. A translation hypothesis $h \in \mathcal{E}$ corresponds

[1] Standardly, the term forced decoding is used to describe the search for only those derivations that exactly produce the reference translation. Our use of this terminology deviates from the standard in two respects: First, we do not require exact reachability of the reference, but only a BOW match. Second, our constraint on the decoder is not strict, but only applies with high probability.

to a sequence of nodes $S \subseteq V$ connected via hyperedges $e$ ending in goal node $g$. Each edge $e$ is associated with a synchronous grammar rule $r(e)$, and corresponding feature values $\mathbf{\Phi}(r(e))$. The weight of hyperedge $e$ is defined as $\mathcal{W}(e; \mathbf{w}) = \mathbf{w}^T \mathbf{\Phi}(r(e))$.

The quantity in (1) is efficiently computed using dynamic programming under the proper semiring. A commutative semiring $K$ is a tuple $\langle \mathbb{K}, \bigoplus, \bigotimes, \bar{0}, \bar{1} \rangle$, of a set $\mathbb{K}$, an associative and commutative addition operator $\bigoplus$, an associative multiplication operator $\bigotimes$, and their "neutral" elements $\bar{0}$ and $\bar{1}$, respectively (Dyer, 2010). The Inside algorithm over the topologically sorted, acyclic hypergraph $\mathcal{H}$ under the tropical $\langle \mathbb{R}, \max, \times, -\infty, 0 \rangle$ semiring (Goodman, 1999; Mohri, 2009) computes the inside score $\alpha$ of the Viterbi hypothesis, i.e. the weight of its sequence of nodes ending in goal node $g$:

$$\begin{aligned} \arg\max_{h \in \mathcal{E}_q} P(h|q) &\equiv \alpha(g) \\ &= \bigoplus_{h \in \mathcal{H}_q} \bigotimes_{e \in h} \mathcal{W}(e; \mathbf{w}_{smt}), \end{aligned}$$

where $\mathcal{W}(e; \mathbf{w}_{smt}) = \mathbf{w}_{smt}^T \mathbf{\Phi}_{smt}(r(e))$ assigns weights given parameters and features of the translation model.

For Bag-of-Words Forced Decoding, we extend $\mathcal{W}$ with another set of parameters $\mathbf{w}_{ir}$ for local IR features $\mathbf{\Phi}_{ir}$:

$$\begin{aligned} \arg\max_{h \in \mathcal{E}_q} P(h|q,d) &\equiv \alpha(g) \\ &= \bigoplus_{h \in \mathcal{H}_q} \bigotimes_{e \in h} \mathcal{W}'(e, d; \mathbf{w}_{smt}, \mathbf{w}_{ir}), \end{aligned} \tag{4}$$

with $\mathcal{W}'(e, d; \mathbf{w}_{smt}, \mathbf{w}_{ir}) = \mathbf{w}_{smt}^T \mathbf{\Phi}_{smt}(r(e)) + \mathbf{w}_{ir}^T \mathbf{\Phi}_{ir}(r(e), d)$. Note that $\mathbf{\Phi}_{ir}$ depends on both translation rule $r(e)$ and document $d$, while $\mathbf{\Phi}_{smt}$ solely depends on source and target side of $r(e)$.

**Decomposable Retrieval Features.** We use sparse, lexicalized, real-valuead IR features that relate derivations $h$ to document $d$ using *Okapi bm25 term weights* (Robertson and Zaragoza, 2009):

$$bm25(t,d) = rsj(t, \mathcal{C}) \cdot tf_{bm25}(t,d),$$

where $rsj(t, \mathcal{C}) = log\left(\frac{|\mathcal{C}| - df(t,\mathcal{C}) + 0.5}{df(t,\mathcal{C}) + 0.5}\right)$ is a constant term weight approximated on document frequencies for collection $\mathcal{C}$, and $tf_{bm25}(t,d) = tf(t,d)/(k_1((1-b) + b\frac{dl}{avdl}) + tf(t,d))$ a saturated term frequency weight of term $t$ in document $d$, taking into account (average) document lengths $dl$ and $avdl$[2]. We fire the Okapi $bm25$ term weight for each derivation term $t \in h$ w.r.t. document $d$ in collection $\mathcal{C}$. The sum of feature values for all derivation terms $t_i \in h$ equals the regular $BM25$ score $BM25(h,d) = \sum_{t \in h} bm25(t,d)$. Weights $\mathbf{w}_{ir}$ for this type of features are interpretable as additional, general term weights.

Additionally, we report experiments using sparse alignment features that fire an indicator for each alignment, insertion, or deletion of words in source and target. They allow the model to adapt lexical choice and dropping of function words for retrieval.

**Default Retrieval Weights & Self-Translation.** To enforce a ranking over documents, we define an *IR default weight* $v$, $\mathbf{w}_{ir} = \mathbf{1}v$. Intuitively, $v$ controls the model's disposition to diverge from the SMT Viterbi path. If IR features fire in other regions of the search space than the SMT Viterbi path, this weight compensates for the loss incurred for not producing the Viterbi hypothesis. Furthermore, the default weight allows the model to generalize to unseen data: If an unknown query word, for example a named entity, causes an IR feature to fire at test time, the decoder will simply *pass through* the source word to any derivation, and the IR feature can contribute to the retrieval score with $v > 0$.

**Multi-Sentence Queries.** Specialized retrieval tasks such as patent prior art search may exhibit long, coherent search queries that contain multiple sentences. If multiple sentences of query $q = (s_1, \ldots, s_m)$ are processed independently, we need to combine the sentence-wise rankings to obtain a final ranking. We model this task from a product of experts perspective (Hinton, 2002) and multiply scores $score(\cdot, d)$ of document $d$ in all $m$ sentence rankings, re-sorting the final output. If $d$ is not in the top-$k$ ranking of a sentence, we take the minimum score of that top-$k$ ranking as a smoothing value to prevent the product to become zero.

[2] bm25 parameters were fixed at $k_1 = 1.2$ and $b = 0.75$

**Implementation and Complexity Analysis.**[3] We implemented the above model on top of the hierarchical phrase-based decoder `cdec` (Dyer et al., 2010), but there are no limitations for applying this approach to phrase-based systems (Koehn et al., 2007). Procedurally, after `cdec` yields the translation forest, we compute the overlap of IR feature activations between edges in the forest and the document candidates. The Inside algorithm is only carried out for documents that activate at least one IR feature in the search space. For documents with no activation we can skip the computation of scores and assign the SMT Viterbi score, which constitutes a lower bound on the model score.

For a single query $q$, forced decoding requires a single pass over the topologically sorted search space to find IR feature activations along hyperedges, yielding a complexity of $O(|V| + |E|)$. The dynamic programming procedure that computes a score for a document requires another pass over the forest evaluating the extended edge weight (4) for every edge $e \in E$, where the dot product for translation features is already precomputed by `cdec` , and the retrieval part depends on the number of active IR features, $\omega := |\mathbf{\Phi}_{ir}(r(e), d)|$. Overall complexity for a single query and all documents $d \in \mathcal{C}$ is thus

$$O\Big(|V| + |E| + \big(|V| + |E| \cdot \omega\big) \cdot |\mathcal{C}|\Big).$$

As noted above, we can reduce the quantity $|\mathcal{C}|$ by checking if a document candidate shares any IR features with the search space and avoid superfluous executions of the Inside algorithm. In our experiments on Wikipedia data, we found that this check reduces $|\mathcal{C}|$ to about $64\%$ of its original size. This pre-filtering is similar to the coarse query approach of Dong et al. (2014), who score only documents that contain at least one term in the query lattice. We further reduce runtime of the inference procedure by using approximate decoding. We experimented with using a beam search approach to limit the number of weight evaluations in (4) for incoming edges at each node. The $\max$ operation of the tropical semiring is discontinued once the number of considered incoming edges at a node exceeds the size of the beam.

## 4 Learning to Decode for Retrieval

We now turn to the problem of learning parameter weights for the BOW-FD model. The objective is to prefer a relevant document $d^+$ over an irrelevant one $d^-$ by assigning a higher score to $d^+$ than to $d^-$,

$$score(q, d^+; \mathbf{w}) > score(q, d^-; \mathbf{w}).$$

We sample a set of preference pairs

$$\mathcal{P} = \{(d^+, d^-) | rl(d^+, q) > rl(d^-, q)\}$$

from relevance-annotated data where $rl(d, q)$ indicates the relevance level of a document given query. Furthermore, we require the difference of scores to satisfy a certain margin:

$$score(q, d^+; \mathbf{w}) > score(q, d^-; \mathbf{w}) + \Delta,$$

where the margin is defined as

$$\Delta = rl(d^+, q) - rl(d^-, q).$$

Our final objective is a margin-rescaled hinge-loss

$$L(\mathcal{P}) = \sum_{d^+, d^- \in \mathcal{P}} \big[score(q, d^-; \mathbf{w}) - score(q, d^+; \mathbf{w}) + \Delta\big]_+$$

where $[\cdot]_+ = max(0, \cdot)$.

We use stochastic (sub)gradient descent optimization using the *Adadelta* (Zeiler, 2012) update rule. Adadelta does not require manual tuning of a global learning rate and requires only two hyperparameters that have shown to be quite robust to changes: the sliding window decay rate $\rho = 0.95$ and a constant $\epsilon = 10^{-6}$ were set to the default parameters given in the original paper. We furthermore use the distributed learning technique of *Iterative Parameter Mixing* (McDonald et al., 2010), where multiple models on several shards of the training data are trained in parallel and parameters are averaged after each epoch. We perform incremental optimization using a *cyclic order* of the data sequence (Bertsekas, 2011), that is, the learner steps through a fixed sequence of pairs, query by query, and relevant document by relevant document, without randomization after epochs. This allows us to cache consecutive query search spaces and feature vectors for relevant documents. Regularization is done by early stopping where the best iteration is found on a held-out development set.

[3] The complexity of the construction of the translation forest including the language model is common to BOW-FD and the other baselines and thus not included in the following analysis.

| | Wikipedia | | | patents | | |
|---|---|---|---|---|---|---|
| | MAP | NDCG | PRES | MAP | NDCG | PRES |
| DT | .3678 | .5691 | .7219 | .2554 | .5397 | .5680 |
| PSQ | .3642 | .5671 | .7165 | .2659 | .5508 | .5851 |
| BOW-FD | $^*$.3880 | $^*$.5911 | $^*$.7417 | $^*$.2825 | $^*$.5721 | $^*$.6072 |
| BOW-FD+LTR | $^\dagger$.3913 | $^\dagger$.5962 | $^\dagger$**.7543** | $^\dagger$.2870 | $^\dagger$.5807 | $^\dagger$**.6260** |
| BOW-FD+LEX+LTR | $^\dagger$**.3919** | $^\dagger$**.5963** | $^\dagger$.7528 | $^\dagger$**.2883** | $^\dagger$**.5819** | $^\dagger$.6251 |

Table 1: Retrieval results of baseline systems and BOW-FD with default weight $v = 1.6$ for Wikipedia and $v = 0.8$ for patents, respectively. Baseline and BOW-FD models use the same SMT system. Significant differences at $p = 10^{-4}$ with respect to baselines are indicated with $^*$. Significant differences at $p = 10^{-6}$ of learning-to-rank-based models (LTR) with respect to BOW-FD are indicated with $^\dagger$.

| | Wikipedia | | | patents | | |
|---|---|---|---|---|---|---|
| | MAP | NDCG | PRES | MAP | NDCG | PRES |
| DT | $.3347^{(-.03)}$ | $.5368^{(-.03)}$ | $.6970^{(-.03)}$ | $.2315^{(-.02)}$ | $.5105^{(-.03)}$ | $.5420^{(-.03)}$ |
| PSQ | $.3464^{(-.02)}$ | $.5483^{(-.02)}$ | $.7006^{(-.02)}$ | $.2460^{(-.02)}$ | $.5290^{(-.02)}$ | $.5672^{(-.02)}$ |
| BOW-FD | $.3218^{(-.07)}$ | $.5315^{(-.06)}$ | $.7220^{(-.02)}$ | $.1651^{(-.12)}$ | $.4185^{(-.15)}$ | $.4959^{(-.11)}$ |

Table 2: SMT-based CLIR models without a language model. Numbers in superscripts denote the absolute loss with respect to equivalent systems in Table 1.

## 5 Evaluation

**Data and Systems.** We conducted experiments on two large-scale CLIR tasks, namely German-English Wikipedia cross-lingual article retrieval[4] (Bai et al., 2010; Schamoni et al., 2014), and patent prior art search with Japanese-English patent abstracts[5] (Fujii et al., 2009; Guo and Gomes, 2009; Sokolov et al., 2013; Schamoni et al., 2014), comparing retrieval performance of BOW-FD against the state-of-the-art SMT-based CLIR baselines of *Direct Translation* (DT) and cross-lingual *Probabilistic Structured Queries* (PSQ) (Ture et al., 2012a; Ture et al., 2012b). The SMT models, as well as baseline evaluation scores were taken from (Schamoni et al., 2014).

We present results for BOW-FD using a default weight $v$ optimized on the development sets, and for models with parameters trained using pairwise learning-to-rank. We compute MAP, NDCG (Manning et al., 2008) and PRES (Magdy and Jones, 2010) scores on the top 1,000 returned documents to provide an extensive evaluation across precision-, and recall-oriented measures. Differences in evaluation scores between two systems were tested for statistical significance using paired randomization tests (Smucker et al., 2007). Significance levels are either indicated as superscripts, or provided in the captions of the respective tables.

Baseline SMT systems and BOW-FD share the hierarchical phrase-based SMT systems built with `cdec` (Dyer et al., 2010). For German-English cross-lingual article retrieval on Wikipedia, we built a system analogously to Schamoni et al. (2014) from parallel training data (over 104M words) consisting of the Europarl corpus (Koehn, 2005) in version 7, the News Commentary corpus, and the Common Crawl corpus (Smith et al., 2013). Word alignments were created with `fast_align` (Dyer et al., 2013). The 4-gram language model was trained with the KenLM toolkit (Heafield, 2011) on the English side of the training data and the English Wikipedia articles. Language model scores are added to the search spaces using the cube pruning algorithm (Huang and Chiang, 2007) with $poplimit = 200$. SMT Model parameters were optimized using MIRA (Chiang et al., 2008) on the WMT2011 news test set (3003

[4] `http://www.cl.uni-heidelberg.de/statnlpgroup/wikiclir/`

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

sentences). The parameters for the baseline PSQ model were found on a development set consisting of 10,000 German queries using 1,000-best lists: interpolation parameter $\lambda = 0.4$, lower threshold $L = 0$, and cumulative threshold $C = 1$.

For the task of Japanese-English patent prior-art search, we use a system analog to Sokolov et al. (2013) and Schamoni et al. (2014). Its SMT features are trained on 1.8M parallel sentences of NTCIR-7 data (Fujii et al., 2008) and weights were tuned on the NTCIR-8 test collection (2,000 sentences) using MIRA (Chiang et al., 2008). A 5-gram language model on the English side of the training data was trained with the KenLM toolkit (Heafield, 2011). The system uses a cube pruning poplimit of 30. Parameters for the baseline PSQ model were found on a development set of 2,000 patent abstract queries and set to $n$-best list size = $1000$, $\lambda = 1.0$, $L = 0.005$, $C = 0.95$

**Experimental Results.** We first find a default weight $v$ using grid search within $v = [0, 3]$ and $v = [0, 2]$ on the development sets for Wikipedia and patents, respectively. $v$ controls the balance between the retrieval and translation features and with larger $v$, the model is more likely to produce query derivations diverging from the SMT 1-best translation. For Wikipedia, we sample 1,000 out of 10,000 queries to reduce the time of the grid search. For patents we use the full development set of 2,000 queries with 8,381 sentences. We combine rankings for single-sentence queries from multi-sentence patent abstracts using the product method as previously described. Well performing values were found at $v = 1.6$ for Wikipedia, and $v = 0.8$ for patents, respectively.

Table 1 shows test set performance of DT and PSQ baselines versus BOW-FD. Scores for DT and PSQ are as reported in Schamoni et al. (2014). We observe that BOW-FD significantly outperforms both baselines by over 2 points on Wikipedia and patents under all three evaluation measures. While the cube pruning poplimit was set to 200 for the Wikipedia experiments, it is set to 30 for patents. This may reduce the diversity of the search space considerably. Increasing the poplimit from 30 to 200 yielded another significant gain (MAP=0.2893, NDCG=0.5807, PRES=0.6172) on this dataset.

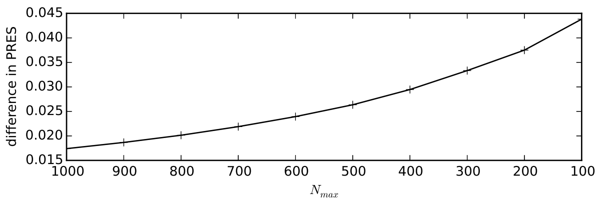

Figure 1: Difference in PRES scores on the Wikipedia development set as a function of PRES's $N_{max}$ parameter between BOW-FD +LM and -LM systems.

**Learning-to-rank results.** We learned the weights of the BOW-FD model starting from IR default weights optimized by grid search, and from SMT feature weights "pre-trained" on parallel data. We furthermore found improvements over BOW-FD in precision-oriented metrics (MAP and NDCG) by freezing SMT weights. Table 1 shows that BOW-FD+(LEX+)LTR models significantly outperform BOW-FD on both data sets, with the largest improvement for PRES. Differences between models with and without lexical alignment features are not statistically significant. We conjecture that LTR models mostly optimize recall rather than precision, i.e. placing more relevant document in the ranking. This is supported by the fact that BOW-FD+LTR retrieves 70.1% of the relevant documents in the test set, compared to 68.0% by BOW-FD, while Mean Reciprocal Rank (MRR) hardly differs (0.7344 vs. 0.7332). An experiment with no pre-trained SMT or default IR weights, performed worse, indicating the importance of translation-benign search spaces and IR default weights for generalization to unseen terms.

**Importance of Language Model for Retrieval.** Liu et al. (2012) and Dong et al. (2014) claim that computationally expensive SMT feature functions such as language models have only minor impact on CLIR performance of SMT-based models. We found that such context-sensitive information present in single 1-best query translations (DT), weighted translation alternatives from the $n$-best list (PSQ), and forced decoding in a "translation-benign" search space (BOW- FD) is crucial for retrieval performance in the experiments reported this paper. In order to investigate the question of the importance of context-sensitive information such as language model scores for retrieval we conducted an experiment in which the language model information is removed from all three SMT-based models. For the PSQ models, we also set the parameter $\lambda$ to 1.0 to disable interpolation with the context-free lexical translation table (Ture et al., 2012a). Table 2 shows that retrieval performance drops significantly for all models. The drop in performance for the two baseline models is comparable on both data sets. Removing the language model for BOW-FD hurts performance the most (with an average drop of 6 points in MAP and NDCG scores for Wikipedia, and over 11 points in all measures for patents). However, scores for recall-oriented PRES on Wikipedia remains relatively stable for BOW-FD with and without a language model. A closer analysis on the rankings for BOW-FD on Wikipedia shows that the -LM model returns 1,589 (out of 86,994) relevant documents less than the +LM model. However, only 2 documents with relevance level 3, i.e., directly linked cross-lingual "mates", were no longer retrieved, suggesting that excluding the language model from the system mostly affects the retrieval of "non-mates", i.e. documents that are linked by, or link to the cross-lingual mate. We explain this behavior as follows: Cross-lingual mates are likely to contain words that are close to an adequate query translation, since they constitute the beginning of a Wikipedia article with the same topic as the query. Derivations generated for these documents are such that both translation model features (with or without the LM) *and* retrieval features agree on a path close to the SMT Viterbi translation. In contrast, other relevant documents require more non-standard lexical choices that are harder to achieve in a +LM search space, since the strong weight on the language model, plus a language model-driven pruning technique, strongly favor lexical choices that agree with the language model's concept of fluency. In a -LM search space, disfluent derivations are easily reached by IR feature activations whose default weight is much larger in relation to the remaining SMT features. The use of "glue rules" allowing left-to-right concatenation of partial translations along with loosely extracted synchronous grammar rules give hierarchical MT models large degrees of freedom in producing very disfluent translations in the -LM space. If a language model is not ensuring a more or less "translation-benign" search space, the "reachability" of terms in irrelevant documents is increased causing them to interfere with the ranking of relevant documents that may be closer translations of the query. This behavior immediately affects precision-oriented scores such as MAP and NDCG, while PRES is only affected if its recall cutoff parameter, $N_{max}$, is lowered, as shown in Figure 1.

The major drop in performance for patent data may be explained with the way multiple sentence queries are evaluated: A language model limits diversity of translation options for multiple sentences. Without a language model, the sets of documents retrieved by each sentence are almost disjoint, i.e. the sentences do not agree on a common set of documents.

## 6 Conclusion

In this paper, we presented an approach to CLIR that shifts the focus from retrieval to translation by forcing a standard SMT decoder to produce a bag-of-words representation of the document repository. This is done by joint optimization of a linear model including both translation and retrieval features under a ranking objective. Highly weighted term-match features are then used to find a decoding path that gives highest score to the document that is optimal with respect to both relevance and translational adequacy. We showed in a large-scale evaluation on cross-lingual retrieval tasks in the domains of patents and Wikipedia pages that our approach significantly outperforms direct translation and Probabilistic Structured Query approaches under a variety of evaluation metrics. Furthermore, we investigated the role of context-sensitive information such as language model scores in retrieval. In contrast to previous claims about the minor impact of language models in retrieval performance in SMT-based CLIR, we found significant drops in MAP and NDCG across all models when removing language model information. This confirms the dual role of the language model to ensure fluency and to select the proper translation terms in the context of the neighboring target terms. The latter role of the language model makes it an indispensable ingredient of any SMT-based CLIR approach.

Open questions in our work regard further improvements in efficiency of retrieval. So far we could achieve substantial reductions in retrieval complexity by pre-filtering based on coarse term matches. The inherent complexity of SMT decoding is less of a problem in offline applications such as translation retrieval (Dong et al., 2014), but it becomes prohibitive in online applications such as cross-lingual web search. In future work, we would like to address efficiency, e.g. by investigating the possibility of incorporating an inverted index into online applications of forced decoding.

## Acknowledgments

This research was supported in part by DFG grant RI-2221/1-2 “Weakly Supervised Learning of Cross-Lingual Systems”.

## References

Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. 2010. Learning to rank with (a lot of) word features. *Information Retrieval*, 13(3):291–314.

Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’99)*, Berkeley, CA.

Dimitri P. Bertsekas. 2011. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors, *Optimization for Machine Learning*. MIT Press.

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP’08)*, Honolulu, Hawaii.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Jeffrey Chin, Maureen Heymans, Alexandre Kojoukhov, Jocelyn Lin, and Hui Tan. 2008. Cross-language information retrieval. Patent Application. US 2008/0288474 A1.

Kareem Darwish and Douglas W. Oard. 2003. Probabilistic structured query methods. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR’03)*, Toronto, Canada.

Meiping Dong, Yong Cheng, Yang Liu, Jia Xu, and Maosong Sun. 2014. Query lattice for translation retrieval. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING’14)*, Dublin, Ireland.

Doug Downey, Susan Dumais, Dan Liebling, and Eric Horvitz. 2008. Understanding the relationship between searchers’ queries and information goals. In *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM’08)*, Napa Valley, California.

Christopher Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL-10 System Demonstrations (ACL’10)*, Uppsala, Sweden.

Christopher Dyer, Victor Chahuneau, and Noah Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies (NAACL-HLT’13)*, Atlanta, Georgia.

Christopher Dyer. 2010. *A Formal Model of Ambiguity and Its Applications in Machine Translation*. Ph.D. thesis, University of Maryland, College Park, Maryland.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2008. Overview of the patent translation task at the NTCIR-7 workshop. In *Proceedings of the 7th NII Testbeds and Community for Information access Research Workshop (NTCIR-7’08)*, Tokyo, Japan.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2009. Evaluating effects of machine translation accuracy on cross-lingual patent retrieval. In *Proceedings of the 32rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR’09)*, Boston, Massachusetts.

Giorgio Gallo, Giustino Longo, Stefano Pallottino, and Sang Nguyen. 1993. Directed hypergraphs and applications. *Discrete Applied Mathematics – Special issue: combinatorial structures and algorithms*, 42(2-3):177–201.

Joshua Goodman. 1999. Semiring parsing. *Computational Linguistics*, 25(4):573–605.

Yunsong Guo and Carla Gomes. 2009. Ranking structured documents: A large margin based approach for patent prior art search. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI’09)*, Pasadena, CA.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation (WMT'11)*, Edinburgh, UK.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

Liang Huang and David Chiang. 2005. Better k-best parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology (IWPT'05)*, Vancouver, Canada.

Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, Prague, Czech Republic.

Dan Klein and Christopher D. Manning. 2001. Parsing and hypergraphs. In *Proceedings of the Seventh International Workshop on Parsing Technologies (IWPT'01)*, Beijing, China.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL-07 2007 Demo and Poster Sessions (ACL'07)*, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, Phuket, Thailand.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, New York, 1st edition.

Chunyang Liu, Qi Liu, Yang Liu, and Maosong Sun. 2012. Thutr: A translation retrieval system. In *Proceedings of COLING'12: Demonstration Papers*, Bombay, India.

Walid Magdy and Gareth Jones Jones. 2010. PRES: A score metric for evaluating recall-oriented information retrieval applications. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*, Geneva, Switzerland.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, New York.

Ryan McDonald, Keith Hall, and Gideon Mann. 2010. Distributed training strategies for the structured perceptron. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies (NAACL-HLT'10)*, Los Angeles, California.

Mehryar Mohri. 2009. Weighted automata algorithms. In *Handbook of weighted automata*, pages 213–254. Springer Berlin Heidelberg.

Vassilina Nikoulina, Bogomil Kovachev, Nikolaos Lagos, and Christof Monz. 2012. Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, Avignon, France.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Shigehiko Schamoni, Felix Hieber, Artem Sokolov, and Stefan Riezler. 2014. Learning translational and knowledge-based similarities from relevance rankings for cross-language retrieval. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, Baltimore, USA.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, Sofia, Bulgaria.

Mark D. Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the 16th ACM conference on Conference on Information and Knowledge Management (CIKM'07)*, New York, New York.

Artem Sokolov, Laura Jehl, Felix Hieber, and Stefan Riezler. 2013. Boosting cross-language retrieval by learning bilingual phrase associations from relevance rankings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, Seattle, WA.

Artem Sokolov, Felix Hieber, and Stefan Riezler. 2014. Learning to translate queries for clir. In *Proceedings of the 37th Annual ACM SIGIR Conference (SIGIR'14)*, Gold Coast, Australia.

Ferhan Ture, Jimmy Lin, and Douglas W. Oard. 2012a. Combining statistical translation techniques for cross-language information retrieval. In *Proceedings of the International Conference on Computational Linguistics (COLING'12)*, Mumbai, India.

Ferhan Ture, Jimmy Lin, and Douglas W. Oard. 2012b. 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'12)*, Portland, Oregon.

Jinxi Xu, Ralph Weischedel, and Chanh Nguyen. 2001. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'01)*, New Orleans, Louisiana.

Heng Yu, Liang Huang, Haitao Mi, and Kai Zhao. 2013. Max-violation perceptron and forced decoding for scalable mt training. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, Seattle, Washington.

Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. *Computing Research Repository (CoRR'2012)*, abs/1212.5701.

Kai Zhao, Liang Huang, Haitao Mi, and Abe Ittycheriah. 2014. Hierarchical mt training using max-violation perceptron. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, Baltimore, Maryland.