Cross-lingual Slot Filling from Comparable Corpora

Matthew Snover, Xiang Li, Wen-Pin Lin, Zheng Chen, Suzanne Tamang,
Mingmin Ge, Adam Lee, Qi Li, Hao Li, Sam Anzaroot, Heng Ji
Computer Science Department
Queens College and Graduate Center
City University of New York
New York, NY 11367, USA
msnover@qc.cuny.edu, hengji@cs.qc.cuny.edu

Abstract

This paper introduces a new task of crosslingual slot filling which aims to discover attributes for entity queries from crosslingual comparable corpora and then present answers in a desired language. It is a very challenging task which suffers from both information extraction and machine translation errors. In this paper we analyze the types of errors produced by five different baseline approaches, and present a novel supervised rescoring based validation approach to incorporate global evidence from very large bilingual comparable corpora. Without using any additional labeled data this new approach obtained 38.5% relative improvement in Precision and 86.7% relative improvement in Recall over several state-of-the-art approaches. The ultimate system outperformed monolingual slot filling pipelines built on much larger monolingual corpora.

1 Introduction

The slot filling task at NIST TAC Knowledge Base Population (KBP) track (Ji et al., 2010) is a relatively new and popular task with the goal of automatically building profiles of entities from large amounts of unstructured data, and using these profiles to populate an existing knowledge base. These profiles consist of numerous slots such as "*title*", "*parents*" for persons and "*top-employees*" for organizations. A variety of approaches have been proposed to address both tasks with considerable success; nevertheless, all of the KBP tasks so far have been limited to monolingual processing. However, as the shrinking fraction of the world's Web pages are written in English, many slot fills can only be discovered from comparable documents in foreign languages. By comparable corpora we mean texts that are about similar topics, but are not in general translations of each other. These corpora are naturally available, for example, many news agencies release multi-lingual news articles on the same day. In this paper we propose a new and more challenging crosslingual slot filling task, to find information for any English query from crosslingual comparable corpora, and then present its profile in English.

We developed complementary baseline approaches which combine two difficult problems: information extraction (IE) and machine translation (MT). In this paper we conduct detailed error analysis to understand how we can exploit comparable corpora to construct more complete and accurate profiles.

Many correct answers extracted from our baselines will be reported multiple times in any external large collection of comparable documents. We can thus take advantage of such information redundancy to rescore candidate answers. To choose the best answers we consult large comparable corpora and corresponding IE results. We prefer those answers which frequently appear together with the query in certain IE contexts, including co-occurring names, coreference links, relations and events. For example, we prefer "*South Korea*" instead of "*New York Stock Exchange*" as the "*per:employee_of*" answer for "*Roh Moo-hyun*" using global evidence from employment relation extraction. Such global knowledge from comparable corpora

Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 110–119,
49th Annual Meeting of the Association for Computational Linguistics,
Portland, Oregon, 24 June 2011. ©*2011 Association for Computational Linguistics*

provides substantial improvement over each individual baseline system and even state-of-the-art monolingual slot filling systems. Compared to previous methods of exploiting comparable corpora, our approach is novel in multiple aspects because it exploits knowledge from: (1) both local and global statistics; (2) both languages; and (3) both shallow and deep analysis.

2 Related Work

Sudo et al. (2004) found that for a crosslingual single-document IE task, source language extraction and fact translation performed notably better than machine translation and target language extraction. We observed the same results. In addition we also demonstrate that these two approaches are complementary and can be used to boost each other's results in a statistical rescoring model with global evidence from large comparable corpora.

Hakkani-Tur et al. (2007) described a filtering mechanism using two crosslingual IE systems for improving crosslingual document retrieval. Many previous validation methods for crosslingual QA, such as those organized by Cross Language Evaluation Forum (Vallin et al., 2005), focused on local information which involves only the query and answer (e.g. (Kwork and Deng, 2006)), keyword translation (e.g. (Mitamura et al., 2006)) and surface patterns (e.g. (Soubbotin and Soubbotin, 2001)). Some global validation approaches considered information redundancy based on shallow statistics including co-occurrence, density score and mutual information (Clarke et al., 2001; Magnini et al., 2001; Lee et al., 2008), deeper knowledge from dependency parsing (e.g. (Shen et al., 2006)) or logic reasoning (e.g. (Harabagiu et al., 2005)). However, all of these approaches made limited efforts at disambiguating entities in queries and limited use of fact extraction in answer search and validation.

Several recent IE studies have stressed the benefits of using information redundancy on estimating the correctness of the IE output (Downey et al., 2005; Yangarber, 2006; Patwardhan and Riloff, 2009; Ji and Grishman, 2008). Some recent research used comparable corpora to re-score name transliterations (Sproat et al., 2006; Klementiev and Roth, 2006) or mine new word translations (Fung and Yee, 1998; Rapp, 1999; Shao and Ng, 2004; Tao and Zhai, 2005; Hassan et al., 2007; Udupa et al., 2009; Ji, 2009). To the best of our knowledge, this is the first work on mining facts from comparable corpora for answer validation in a new crosslingual entity profiling task.

3 Experimental Setup

3.1 Task Definition

The goal of the KBP slot filling task is to extract facts from a large source corpus regarding certain attributes ("*slots*") of an entity, which may be a person or organization, and use these facts to augment an existing knowledge base (KB). Along with each slot answer, the system must provide the ID of a document which supports the correctness of this answer. KBP 2010 (Ji et al., 2010) defines 26 types of attributes for persons (such as the age, birthplace, spouse, children, job title, and employing organization) and 16 types of attributes for organizations (such as the top employees, the founder, the year founded, the headquarters location, and the subsidiaries).

The new problem we define in this paper is an extension of this task to a crosslingual paradigm. Given a query in a target language t and a collection of documents in a source language s, a system must extract slot answers about the query and present the answers in t. In this paper we examine a specific setting of s=Chinese and t=English.

To score crosslingual slot filling, we pool all the system responses and group equivalent answers into equivalence classes. Each system response is rated as correct, wrong, inexact or redundant. Given these judgments, we calculate the precision, recall and F-measure of each system, crediting only correct answers.

3.2 Data and Query Selection

We use the comparable corpora of English TDT5 (278,358 documents) and Chinese TDT5

(56,424 documents) as our source collection.

For query selection, we collected all the entities from the entire source collection and counted their frequencies. We then selected 50 informative entities (25 persons and 25 organizations) which were located in the middle range of frequency counts. Among the 25 person queries, half are Chinese-specific names, and half are non-Chinese names. The 25 organizations follow a representative distribution according to the entity subtypes defined in NIST Automatic Content Extraction (ACE) program[1].

3.3 Baseline Pipelines

3.3.1 Overview

We employ the following two types of baseline crosslingual slot filling pipelines to process Chinese documents. Figure 1 and Table 1 shows the five system pipelines we have used to conduct our experiments.

Type A Translate Chinese texts into English, and apply English slot filling systems to the translations.

Type B Translate English queries into Chinese, apply Chinese slot filling systems to Chinese texts, and translate answers back to English.

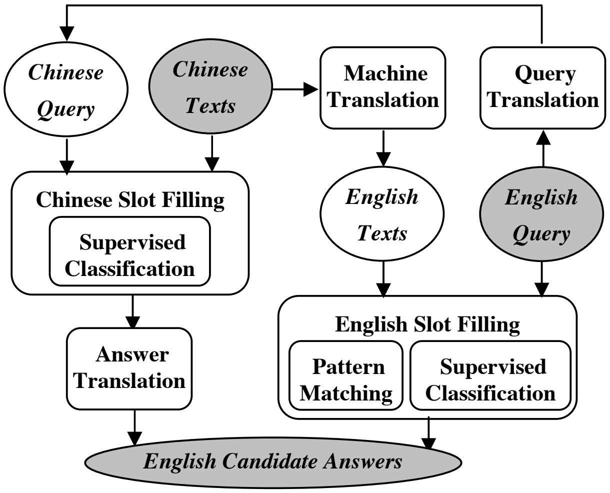

Figure 1: Overview of Baseline Crosslingual Slot Filling Pipelines

Pipeline	Label		Components	Data
Mono-lingual	(1)		English Supervised Classification	English TDT5
	(2)		English Pattern Matching	
Cross-lingual	Type A	(3)	MT+English Supervised Classification	Chinese TDT5
		(4)	MT+English Pattern Matching	
	Type B	(5)	Query Translation +Chinese Supervised Classification +Answer Translation	

Table 1: Monolingual and Crosslingual Baseline Slot Filling Pipelines

3.3.2 Monolingual Slot Filling

We applied a state-of-the-art bilingual slot filling system (Chen et al., 2010) to process bilingual comparable corpora. This baseline system includes a supervised ACE IE pipeline and a bottom-up pattern matching pipeline. The IE pipeline includes relation extraction and event extraction based on maximum entropy models that incorporate diverse lexical, syntactic, semantic and ontological knowledge. The extracted ACE relations and events are then mapped to KBP slot fills. In pattern matching, we extract and rank patterns based on a distant supervision approach (Mintz et al., 2009) that uses entity-attribute pairs from Wikipedia Infoboxes and Freebase (Bollacker et al., 2008). We set a low threshold to include more answer candidates, and then a series of filtering steps to refine and improve the overall pipeline results. The filtering steps include removing answers which have inappropriate entity types or have inappropriate dependency paths to the entities.

3.3.3 Document and Name Translation

We use a statistical, phrase-based MT system (Zens and Ney, 2004) to translate Chinese documents into English for Type A Approaches. The best translation is computed by using a weighted log-linear combination of various statistical models: an n-gram language model, a phrase translation model and a word-based lex-

[1] http://www.itl.nist.gov/iad/mig/tests/ace/

icon model. The latter two models are used in source-to-target and target-to-source directions. The model scaling factors are optimized with respect to the BLEU score similar to (Och, 2003). The training data includes 200 million running words in each language. The total language model training data consists of about 600 million running words.

We applied various name mining approaches from comparable corpora and parallel corpora, as described in (Ji et al., 2009) to extract and translate names in queries and answers in Type B approaches. The accuracy of name translation is about 88%. For those names not covered by these pairs, we relied on Google Translate [2] to obtain results.

4 Analysis of Baseline Pipelines

In this section we analyze the coverage (Section 4.1) and precision (Section 4.2) results of the baseline pipelines. We then illustrate the potential for global validation from comparable corpora through a series of examples.

4.1 Coverage Analysis: Toward Information Fusion

Table 2 summarizes the Precision (P), Recall (R) and F-measure (F) of baseline pipelines and the union of their individual results.

Table 2: Baseline Pipeline Results

System		P	R	F
Mono-lingual	(1)	0.08	0.54	0.15
	(2)	0.02	0.35	0.03
	Union of (1)+(2)	0.03	0.69	0.05
Cross-lingual	(3)	0.04	0.04	0.04
	(4)	0.03	0.25	0.05
	Union of (3)+(4)	0.03	0.26	0.05
	(5)	0.04	0.46	0.08
	Union of (3)+(4)+(5)	0.03	0.56	0.05
Comparable Corpora	Union of (1)+(2)+(3)+ (4)+(5)	0.02	1	0.04

Although crosslingual pipelines used a much smaller corpus than monolingual pipelines, they extracted comparable number of correct answers (66 vs. 81) with a slightly better precision. In fact, the crosslingual pipeline (5) performs even better than monolingual pipeline (2), especially on the employment slots. In particular, 96.35% of the correct answers for Chinese-specific person queries (e.g. "*Tang Jiaxuan*") were extracted from Chinese data. Even for those facts discovered from English data, they are about quite general slots such as "*title*" and "*employee_of*". In contrast, Chinese data covers more diverse biographical slots such as "*family members*" and "*schools_attended*".

Compared to the union of Type A approaches (pipelines (3)+(4)), Pipeline (5) returned many more correct answers with higher precision. The main reason is that Type A approaches suffer from MT errors. For example, MT mistakenly translated the query name "*Celine Dion*" into "*Clinton*" and thus English slot filling components failed to identify any answers. One can hypothesize that slot filling on MT output can be improved by re-training extraction components directly from MT output. However, our experiments of learning patterns from MT output showed negative impact, mainly because MT errors were too diverse to generalize. In other cases even though slot filling produced correct results, MT still failed to translate the answer names correctly. For example, English slot filling successfully found a potential answer for "*org:founded_by*" of the query "*Microsoft*" from the following MT output: "*The third largest of the Microsoft common founder Alan Doss , aged 50, and net assets of US 22 billion.*"; however, the answer string "*Paul Allen*" was mistakenly translated into "*Alan Doss*". MT is not so crucial for "*per:title*" slot because it does not require translation of contexts.

To summarize, 59% of the missing errors were due to text, query or answer translation errors and 20% were due to slot filling errors. Nevertheless, the union of (3)+(4)+(5) still contain more correct answers. These baseline pipelines were developed from a diverse set of algorithms, and typically showed strengths in specific slots.

[2] http://translate.google.com/

In general we can conclude that monolingual and crosslingual pipelines are complementary. Combining the responses from all baseline pipelines, we can get similar number of correct answers compared to one single human annotator.

4.2 Precision Analysis: Toward Global Validation

The spurious errors from baseline crosslingual slot filling pipelines reveal both the shortcomings of the MT system and extraction across languages. Table 3 shows the distribution of spurious errors.

Pipeline	Spurious Errors	Distribution
Type A	Content Translation + Extraction	85%
	Query Translation	13%
	Answer Translation	2%
Type B	Word Segmentation	34%
	Relation Extraction	33%
	Coreference	17%
	Semantic Type	13%
	Slot Type	3%

Table 3: Distribution of Spurious Errors

Table 3 indicates a majority (85%) of spurious errors from Type A pipelines were due to applying monolingual slot filling methods to MT output which preserves Chinese structure.

As demonstrated in previous work (e.g. (Parton and McKeown, 2010; Ji et al., 2009)), we also found that many (14.6%) errors were caused by the low quality of name translation for queries and answers.

For example, "麦克金蒂/*McGinty*" was mistakenly translated into the query name "*Kim Jong-il*", which led to many incorrect answers such as "*The British Royal joint military research institute*" for "*per:employee_of*".

In contrast, the spurious errors from Type B pipelines were more diverse. Chinese IE components severely suffered from word segmentation errors (34%), which were then directly propagated into Chinese document retrieval and slot filling. Many segmentation errors occurred with out-of-vocabulary names, especially person names and nested organization names. For example, the name "姚明宝/*Yao Mingbao*" was mistakenly segmented into two words "姚明/*Yao Ming*" and "宝/*bao*", and thus the document was mistakenly retrieved for the query '*Yao Ming*".

In many cases (33%) Chinese relation and event extraction components failed to capture Chinese-specific structures due to the limited size of training corpora. For example, from the context "应邀担任陈水扁经济顾问的萧万长/*Xiao Wan-chang, who were invited to become the economics consultant for Chen Shui-bian*", Chinese slot filling system mistakenly extracted "*consultant*" as a "*per:title*" answer for the query "*Chen Shui-bian*" using a common pattern "*<query><title>*".

13% of errors were caused due to invalid semantic types for certain slots. For example, many metaphoric titles such as "*tough guy*" don't match the definition of "*per:title*" in the annotation guideline "*employment or membership position*".

5 Global Validation

Based on the above motivations we propose to incorporate global evidence from a very large collection of comparable documents to refine local decisions. The central idea is to over-generate candidate answers from multiple weak baselines to ensure high upper-bound of recall, and then conduct effective global validation to filter spurious errors while keeping good answers in order to enhance precision.

5.1 Supervised Rescoring

Ideally, we want to choose a validation model which can pick out important features in a context wider than that used by baseline pipelines. Merging individual systems to form the union of answers can be effective, but Table 2 shows that simple union of all pipelines produced worse F-measure than the best pipeline.

In this paper we exploit the reranking paradigm, commonly used in information retrieval, to conduct global validation. By modeling the empirical distribution of labeled training data, statistical models are used to identify the strengths and weaknesses (e.g. high and low precision slots) of individual systems, and rescore answers accordingly. Specially, we develop a supervised Maximum Entropy (MaxEnt) based model to rescore the answers from the pipelines, selecting only the highest-scoring answers.

The rescorer was trained (using cross-validation) on varying subsets of the features. The threshold at which an answer is deemed to be true is chosen to maximize the F-Measure on the training set.

5.2 Validation Features

Table 4 describes the validation features used for rescoring, where q is the query, q' the Chinese translation of q, t the slot type, a the candidate answer, a' the Chinese form of a, s the context sentence and d is the context document supporting a.

The feature set benefits from multiple dimensions of crosslingual slot filling. These features were applied to both languages wherever annotation resources were available.

In the KBP slot filling task, slots are often dependent on each other, so we can improve the results by improving the "coherence" of the story (i.e. consistency among all generated answers - query profiles). We use feature *f2* to check whether the same answer was generated for conflicting slots, such as *per:parents* and *per:children.*

Compared to traditional QA tasks, slot filling is a more fine-grained task in which different slots are expected to obtain semantically different answers. Therefore, we explored semantic constraints in both local and global contexts. For example, we utilized bilingual name gazetteers from ACE training corpora, Google n-grams (Ji and Lin, 2009) and the geonames website [3] to encode features *f6*, *f8* and *f9*; The *org:top_members/employees* slot requires a system to distinguish whether a person member/ employee is in the top position, thus we encoded *f10* for this purpose.

The knowledge used in our baseline pipelines is relatively static – it is not updated during the extraction process. Achieving high performance for cross-lingual slot filling requires that we take a broader view, one that looks outside a single document or a single language in order to exploit global knowledge. Fortunately, as more and more large crosslingual comparable corpora are available, we can take advantage of information redundancy to validate answers. The basic intuition is that if a candidate answer a is correct, it should appear together with the query q repeatedly, in different documents, or even in certain coreference links, relations and events.

For example, "*David Kelly - scientist*", and "石原慎太郎/*Shintaro Ishihara* - 知事/*governor*" pairs appear frequently in "*title*" coreference links in both English and Chinese corpora; "*Elizabeth II*" is very often involved in an "*employment*" relation with "*United Kingdom*" in English corpora. On the other hand, some incorrect answers with high global statistics can be filtered out using these constraints. For example, although the query "唐家璇/*Tang Jiaxuan*" appears frequently together with the candidate *per:title* answer "人员/*personnel*", it is linked by few coreference links; in contrast, it's coreferential with the correct title answer "国务委员/*State Council member*" much more frequently.

We processed cross-lingual comparable corpora to extract coreference links, relations and events among mentions (names, nominals and time expressions etc.) and stored them in an external knowledge base. Any pair of <q, a> is then compared to the entries in this knowledge base. We used 157,708 documents from Chinese TDT5 and Gigaword to count Chinese global statistics, and 7,148,446 documents from DARPA GALE MT training corpora to count English global statistics, as shown in features *f12* and *f13.* Fact based global features *f14*, *f15*, *f16* and *f17*, were calculated from 49,359 Chinese and 280,513 English documents (annotated by the bilingual IE system in Section 3.3.2.

6 Experiments

In this section, we examine the overall performance of this method. We then discuss the usefulness of the individual sets of features. In

[3] http://www.geonames.org/statistics/

Characteristics			Description
Scope	Depth	Language	
Global (Cross-system)	Shallow	English	f1: frequency of <*q, a, t*> that appears in all baseline outputs
			f2: number of conflicting slot types in which answer *a* appears in all baseline outputs
Local	Shallow	English	f3: conjunction of *t* and whether *a* is a year answer
			f4: conjunction of *t* and whether *a* includes numbers or letters
	Deep	English	f5: conjunction of place *t* and whether *a* is a country name
			f6: conjunction of *per:origin t* and whether *a* is a nationality
			f7: if *t=per:title*, whether *a* is an acceptable title
			f8: if *t* requires a name answer, whether *a* is a name
			f9: whether *a* has appropriate semantic type
Global (Within-Document)	Deep	English	f10: conjunction of *org:top_members/employees* and whether there is a high-level title in *s*
			f11: conjunction of alternative name and whether *a* is an acronym of *q*
Global (Cross-document in comparable corpora)	Shallow (Statistics)	Chinese	f12: conditional probability of *q/q'* and *a/a'* appear in the same document
		English	f13: conditional probability of *q/q'* and *a/a'* appear in the same sentence
	Deep (Fact-based)	Both	f14: co-occurrence of *q/q'* and *a/a'* appear in coreference links
		English	f15: co-occurrence of *q/q'* and *a/a'* appear in relation/event links
		English	f16: conditional probability of *q/q'* and *a/a'* appear in relation/event links
		English	f17: mutual information of *q/q'* and *a/a'* appear in relation/event links

Table 4: Validation Features for Crosslingual Slot Filling

the following results, the baseline features are always used in addition to any other features.

6.1 Overall Performance

Because of the data scarcity, ten-fold cross-validation, across queries, was used to train and test the system. Quantitative results after combining answers from multiple pipelines are shown in Table 5. We used two basic features, one is the slot type and the other is the entity type of the query (i.e. person or organization). This basic feature set is already successful in improving the precision of the pipelines, although this results in a number of correct answers being discarded as well. By adding the additional validation features described previously, both the f-score and precision of the models are improved. In the case of the cross-lingual pipelines (3+4+5) the number of correct answers chosen is almost doubled while increasing the precision of the output.

6.2 Impact of Global Validation

A comparison of the benefits of global versus local features are shown in Table 6, both of which dramatically improve scores over the baseline features. The global features are universally more beneficial than the local features, although the local features generate results with higher precision at the expense of the number of correct answers returned. The global features are especially useful for pipelines 3+4+5, where the performance using just these features reaches those of using all other features – this does not hold true for the monolingual pipelines however.

Pipelines	F	P	R
Basic Features			
1+2	0.31	0.31	0.30
3+4+5	0.26	0.39	0.20
1+2+3+4+5	0.27	0.29	0.25
Full Features			
1+2	0.37	0.30	0.46
3+4+5	0.36	0.35	0.37
1+2+3+4+5	0.31	0.28	0.35

Table 5: Using Basic Features to Filter Answers

6.3 Impact of Fact-driven Deep Knowledge

The varying benefit of fact-driven cross-document features and statistical cross-document features are shown in Table 7.

Pipelines	F	P	R
Local Features			
1+2	0.34	0.35	0.33
3+4+5	0.29	0.40	0.22
1+2+3+4+5	0.27	0.32	0.24
Global Features			
1+2	0.35	0.30	0.42
3+4+5	0.37	0.36	0.38
1+2+3+4+5	0.33	0.29	0.38

Table 6: The Benefit of Global versus Local Features

While both feature sets are beneficial, the monolingual pipelines (1+2) benefit more from statistical features while the cross-lingual pipelines (3+4+7) benefit slightly more from the fact-based features. Despite this bias, the overall results when the features are used in all pipelines are very close with the fact-based features being slightly more useful overall.

Pipelines	F	P	R
Fact-Based Features			
1+2	0.33	0.27	0.42
3+4+5	0.35	0.43	0.29
1+2+3+4+5	0.30	0.27	0.34
Statistical Features			
1+2	0.37	0.34	0.40
3+4+5	0.34	0.35	0.33
1+2+3+4+5	0.29	0.25	0.34

Table 7: Fact vs. Statistical Cross-Doc Features

Translation features were only beneficial to pipelines 3, 4, and 5, and provided a slight increase in precision from 0.39 to 0.42, but provided no noticeable benefit when used in conjunction with results from pipelines 1 and 2. This is because the answers where translation features would be most useful were already being selected by pipelines 1 and 2 using the baseline features.

6.4 Discussion

The use of any re-scoring, even with baseline features, provides large gains over the union of the baseline pipelines, removing large number of incorrect answers. The use of more sophisticated features provided substantial gains over the baseline features. In particular, global features proved very effective. Further feature engineering to address the remaining errors and the dropped correct answer would likely provide increasing gains in performance.

In addition, two human annotators, independently, conducted the same task on the same data, with a second pass of adjudication. The F-scores of inter-annotator agreement were 52.0% for the first pass and 73.2% for the second pass. This indicates that slot filling remains a challenging task for both systems and human annotators—only one monolingual system exceeded 30% F-score in the KBP2010 evaluation.

7 Conclusion and Future Work

Crosslingual slot filling is a challenging task due to limited performance in two separate areas: information extraction and machine translation. Various methods of combining techniques from these two areas provided weak yet complementary baseline pipelines. We proposed an effective approach to integrate these baselines and enhance their performance using wider and deeper knowledge from comparable corpora. The final system based on cross-lingual comparable corpora outperformed monolingual pipelines on much larger monolingual corpora.

The intuition behind our approach is that over-generation of candidate answers from weak baselines provides a potentially strong recall upper-bound. The remaining enhancement becomes simpler: filtering errors. Our experiments also suggest that our rescoring models tend to over-fit due to small amount of training data. Manual annotation and assessment are quite costly, motivating future work in active learning and semi-supervised learning methods. In addition, we plan to apply our results as feedback to improve MT performance on facts using query and answer-driven language model adaptation. We have demonstrated our approach on English-Chinese pair, but the framework is language-independent; ultimately we would like to extend the task to extracting information from more languages.

Acknowledgments

This work was supported by the U.S. NSF CAREER Award under Grant IIS-0953149 and PSC-CUNY Research Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

K. Bollacker, R. Cook, and P. Tufts. 2008. Freebase: A shared database of structured general human knowledge. In *Proc. National Conference on Artificial Intelligence*.

Zheng Chen, Suzanne Tamang, Adam Lee, Xiang Li, Marissa Passantino, and Heng Ji. 2010. Top-down and bottom-up: A combined approach to slot filling. *Lecture Notes in Computer Science*, 6458:300–309, December.

C. L. A. Clarke, G. V. Cormack, and T.R. Lynam. 2001. Exploiting redundancy in question answering. In *Proc. SIGIR2001*.

Doug Downey, Oren Etzioni, and Stephen Soderland. 2005. A Probabilistic Model of Redundancy in Information Extraction. In *Proc. IJCAI 2005*.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel and comparable texts. In *COLING-ACL*.

Dilek Hakkani-Tur, Heng Ji, and Ralph Grishman. 2007. Using information extraction to improve cross-lingual document retrieval. In *Proc. RANLP workshop on Multi-source, Multilingual Information Extraction and Summarization*.

S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, A. Hickl, and P. Wang. 2005. Employing two question answering systems in trec 2005. In *Proc. TREC2005*.

Ahmed Hassan, Haytham Fahmy, and Hany Hassan. 2007. Improving named entity translation by exploiting comparable and parallel corpora. In *RANLP*.

Heng Ji and Ralph Grishman. 2008. Refining Event Extraction through Cross-Document Inference. In *Proc. of ACL-08: HLT*, pages 254–262.

Heng Ji and Dekang Lin. 2009. Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *Proc. PACLIC2009*.

Heng Ji, Ralph Grishman, Dayne Freitag, Matthias Blume, John Wang, Shahram Khadivi, Richard Zens, and Hermann Ney. 2009. Name translation for distillation. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*.

Heng Ji, Ralph Grishman, Hoa Trang Dang, and Kira Griffitt. 2010. An overview of the tac2010 knowledge base population track. In *Proc. TAC2010*.

Heng Ji. 2009. Mining name translations from comparable corpora by creating bilingual information networks. In *ACL-IJCNLP 2009 workshop on Building and Using Comparable Corpora (BUCC 2009): from Parallel to Non-parallel Corpora*.

Alexandre Klementiev and Dan Roth. 2006. Named entity transliteration and discovery from multilingual comparable corpora. In *HLT-NAACL 2006*.

K.-L. Kwork and P. P. Deng. 2006. Chinese question-answering: Comparing monolingual with english-chinese cross-lingual results. In *Asia Information Retrieval Symposium*.

Cheng-Wei Lee, Yi-Hsun Lee, and Wen-Lian Hsu. 2008. Exploring shallow answer ranking features in cross-lingual and monolingual factoid question answering. *Computational Linguistics and Chinese Language Processing*, 13:1–26, March.

B. Magnini, M. Negri, R. Prevete, and H. Tanev. 2001. Is it the right answer?: Exploiting web redundancy for answer validation. In *Proc. ACL2001*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL-IJCNLP 2009*.

Teruko Mitamura, Mengqiu Wang, Hideki Shima, and Frank Lin. 2006. Keyword translation accuracy and cross-lingual question answering in chinese and japanese. In *EACL 2006 Workshop on MLQA*.

F. J. Och. 2003. Minimum error rate training in statistical machine translaton. In *Proc.ACL2003*.

Kristen Parton and Kathleen McKeown. 2010. Mt error detection for cross-lingual question answering. *Proc. COLING2010*.

Siddharth Patwardhan and Ellen Riloff. 2009. A Unified Model of Phrasal and Sentential Evidence for Information Extraction. In *Proc. EMNLP 2009*.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *ACL 1999*.

Li Shao and Hwee Tou Ng. 2004. Mining new word translations from comparable corpora. In *COLING2004*.

D. Shen, G. Saarbruechen, and D. Klakow. 2006. Exploring correlation of dependency relation paths for answer extraction. In *Proc. ACL2006*.

M. M. Soubbotin and S. M. Soubbotin. 2001. Patterns of potential answer expressions as clues to the right answers. In *Proc. TREC2001*.

Richard Sproat, Tao Tao, and ChengXiang Zhai. 2006. Named entity transliteration with comparable corpora. In *ACL 2006*.

Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. 2004. Cross-lingual information extraction evaluation. In *Proc. COLING2004*.

Tao Tao and Chengxiang Zhai. 2005. Mining comparable bilingual text corpora for cross-language information integration. In *Proc. KDD2005*.

Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2009. Mint: A method for effective and scalable mining of named entity transliterations from large comparable corpora. In *EACL2009*.

Alessandro Vallin, Bernardo Magnini, Danilo Giampiccolo, Lili Aunimo, Christelle Ayache, Petya Osenova, Anselmo Peas, Maaren de Rijke, Bogdan Sacaleanu, Diana Santos, and Richard Sutcliffe. 2005. Overview of the clef 2005 multilingual question answer track. In *Proc. CLEF2005*.

Roman Yangarber. 2006. Verification of Facts across Document Boundaries. In *Proc. International Workshop on Intelligent Information Access*.

Richard Zens and Hermann Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proc. HLT/NAACL 2004*.