# NYU-Fair Isaac-RWTH 07 Chinese to English Entity Translation System

Heng Ji+, Matthias Blume*, Dayne Freitag*, Ralph Grishman+, Shahram Khadivi#, and Richard Zens#

+ New York University * Fair Isaac Corporation

# Rheinisch-Westfälische Technische Hochschule Aachen

**Abstract**

In this report we present the overall architecture of the Chinese to English ACE 2007 EDR and ET (Entity Translation) system developed by NYU, Fair Isaac and RWTH. Central to our research is the idea that global knowledge, which comes from within-document and cross-document coreference for both source and target languages, can help to correct MT results and select the best translation. Our system combines the results of several different entity translation strategies. We incorporated two separate name transliterations techniques to correct MT results. We also combine Chinese EDR and English EDR systems according to different entity types, and encode the difference between two language EDRs to improve ET results. In addition, we incorporated the feedback from entity translation to improve Chinese EDR.

## 1 System Overview

Our Chinese to English Entity Translation (ET) system, developed primarily to aid the GALE evaluations, focuses on processing original Chinese news texts, instead of the texts translated from other languages such as English or Arabic. The ET system is built on top of NYU's English and Chinese EDR systems, RWTH's MT and name transliteration systems, and Fair Isaac's name translation systems.

Sudo et al. (2004) evaluated two possible pipelines to extract and translate entities:

**Pipeline 1.** Translate source language texts into target language, and then run target language EDR.

**Pipeline 2.** Run source language EDR, and then translate extracted entities into the target language.

They demonstrated that Pipeline 2 performs substantially better than Pipeline 1. Based on our experiments on the ET development set we found that this observation holds, assuming that the source EDR system doesn't perform much worse than the target EDR system; and that the loss by source EDR system is much less than MT.

Furthermore we found that if we divide the entities into different types and apply different pipelines and then merge results, the system will be more effective. For our set of components, Pipeline 2 generally performs better than Pipeline 1 on Person, GPE and ORG entities; but worse on Location, Facility, Weapon and Vehicle entities. On the other hand, the bottleneck of Pipeline 1 is the performance of MT, and it's quite costly to develop an entity extraction oriented MT system. So Pipeline 2 is more flexible in the sense that

we can integrate other name transliteration and translation techniques as post-processing to correct the MT output.

The overall architecture of our secondary ET system is presented in Figure 1. The primary system keeps the same structure as the secondary system, except it does not include the MT name transliteration step (Pipeline 2.3 as shown in Figure 1).

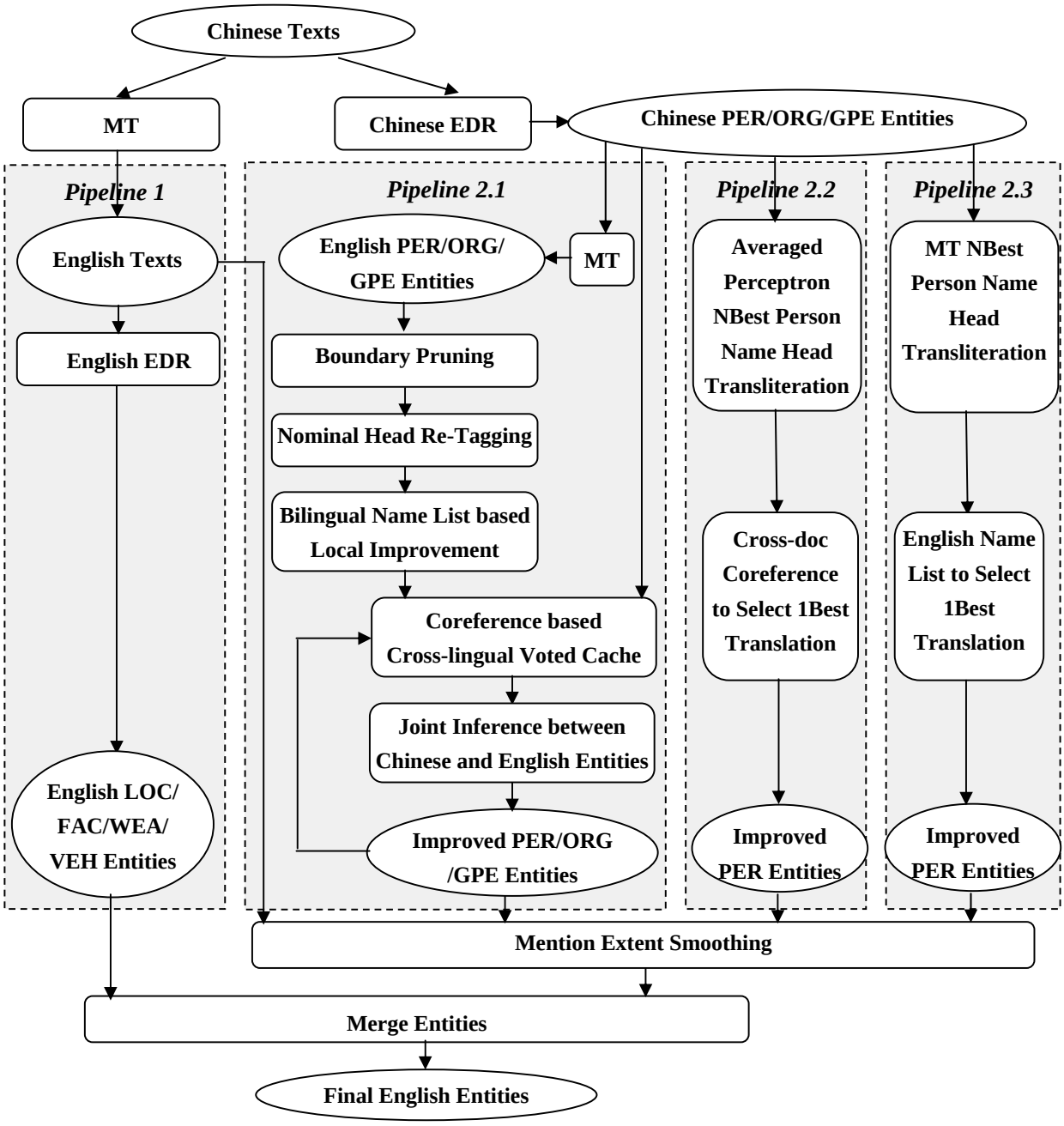


Figure 1. NYU-FI-RWTH Chinese to English ET07 System

In the regular evaluation the primary system required about 4.5 hours using 1 PC, while the secondary system required about 8 hours on 1 PC; In the diagnostic evaluation the primary system required about 3.5 hours using 1 PC, while the secondary system required about 7.2 hours on 1 PC.

## 2 Pipeline1: Target Language EDR based ET

### 2.1 MT System

We use a phrase-based translation system similar to [Zens et al. 2004]. The system memorizes all phrasal translations that have been observed in the training corpus. It computes the best translation using a weighted log-linear combination of various statistical models: an n-gram language model, a phrase translation model and a word-based lexicon model. The latter two models are used in source-to-target and target-to-source directions. Additionally, we use a word penalty and a phrase penalty.

The model scaling factors are optimized on the development corpus with respect to the BLEU score similar to [Och 2003].

Note that the MT system does NOT include a special named entity translation module.

For the Chinese-English ET task we make use of almost all bilingual corpora provided by LDC, which account for about 200 million running words in each language. For language modeling, we use the English part of the bilingual training corpus and in addition some parts of the English GigaWord corpus. The total language model training data consists of about 600 million running words.

### 2.2 English EDR System

The English ACE EDR system [Grishman et al. 2005] includes the following main components:

**Lexical Lookup.** The input English text is divided into sentences and tokenized. Tokens are looked up in a large general English dictionary that provides part-of-speech information and the base form of inflected words.

**Name Tagging**. Named entities are tagged using an HMM trained on the ACE training corpora. The HMM has six states for each name type (Person, GPE, …), as well as a not-a-name state. These six states correspond to the token preceding the name; the single name token (for names with only one token); the first token of the name; an internal token of the name (neither first nor last); the last token of the name; and the token following the name. These multiple states allow the HMM to capture context information and limited information about the internal structure of the name.

**Reference resolution**: Reference resolution first identifies mentions (referring expressions) and then links co-referring expressions. We used a hand-coded rule-based system for ET evaluation.

**ACE Entity Detection and Recognition**. Given the output of reference resolution, ACE entity detection is primarily a task of semantic classification of the co-referring mention groups. Classification is performed differently for entities with and without named mentions. If there is a named mention, the type of the entity is obtained from the name tagger. The subtype is determined using a MaxEnt model whose features are the

individual tokens of the name, trained on the ACE 2005 corpus. The type and subtype of a nominal mention is determined from the head of the mention (with the exception of a small number of hand-coded cases), with the frequencies of the types and subtypes learned from the ACE corpus.

## 3 Pipeline2: Source Language EDR based ET

In pipeline 2 we first apply our Chinese EDR system [Ji and Grishman 2005] to extract Chinese entities, and then translate these entities in isolation into English using the MT system. Then we apply separate post-processing name transliteration/translation techniques to correct the translations.

### 3.1 Chinese EDR

We present the main components in our Chinese EDR system as follows.

**Name Tagging.** We apply a HMM tagger to identify the main ACE entity types: Person, GPE, Organization and Location. The HMM tagger generally follows the Nymble model [Bikel et al, 1997], and operates on the output of a word segmenter from Tsinghua University. We extended the HMM into 14 states based on name structure parsing. We also extract event trigger word lists, a title list and a TIME word list from ACE05 event and TIMEX data as additional features. The training data consists of the Beijing Corpus with Location manually separated from GPE, ACE03, 04, 05 training data and ET07 training data, plus the data generated from ACE05 un-annotated data by semi-supervised learning.

The HMM tagger uses the best-first search to generate NBest hypotheses, and also computes the margin – the difference between the log probabilities of the top two hypotheses. This is used as a rough measure of confidence in the top hypothesis. We then incorporate the results from later components, namely coreference analysis, relation extraction, and event patterns, into statistical re-ranking models to determine a new ranking, and output the new top name hypothesis. Finally we apply a set of post-processing heuristic rules to correct some omissions and systematic errors of the best hypothesis.

**Nominal Mention Tagging.** We trained a SVM based chunker on Chinese Penn TreeBank V5.1 to process the segmented data, and then classify entity types for the nominal heads by list matching. The lists are generated from ACE 03, 04 and 05 training data and ET07 training data, and a small part from Hownet and LDC lists.

**Entity SubType Tagging.** For nominal and pronoun subtype tagging, we count the frequency for each pair of <nominal, frequency>, and assign the subtype with maximum probability to each test mention. If the nominal does not appear in the training data, we assign it the most common subtype in the training corpus. We trained a multi-class tagger to identify name subtypes. The training data consists of entity subtype tagging in ACE 05 training data, and ACE 04 training data with subtypes mapped to 05 subtype set. To generate the subtype of an entity from its mentions, we use the following rule: If an entity includes names, assign its subtype as the first name's subtype, otherwise assign it as the first non-pronoun's subtype.

**Reference Resolution.** The coreference resolver goes through two successive stages. First, high-precision heuristic rules make some positive and negative reference decisions about mention pairs on the basis of

string matching and syntactic analysis. The remaining pairs are assigned confidence values by a combination of maximum entropy models. Since different mention types have different coreference problems, we separate the system into different maximum entropy classifiers for names, nominals, and pronouns. Each classifier operates on a distinct feature set.

### 3.2 Context Independent Mention Translation

For each individual Chinese mention extent/head produced by the EDR system, we use almost the same MT system as described in section 2.1 to translate it in isolation. Here, the only difference is that, as we are translating subsentential units, we do not assume sentence boundaries at the beginning and end of each input segment.

We applied this isolated translation scheme instead of the entity projection based on word alignment because: (1) It produces less alignment noise. Note the word alignments are indirectly derived from phrase alignment, and thus context words often become noise for mention translation; (2) Manual evaluation on a small development set showed that isolated translation obtains (about 14%) better F-measure; (3) ET 07 evaluation doesn't require identifying mention offsets, so isolated translation is more efficient.

### 3.3 Boundary Pruning

We post-process MT mention extent and head translations by correcting the boundaries based on some heuristic, language-specific rules. For example, we remove stop words at the beginning or end of a name, such as ",", "'s", "at", "the", "who", "which", "of", "in", "by", "and", "a", ";", "to", etc. So that we can correct "(科威特)of Kuwait this war" into "Kuwait", "(纳西里耶) in Nasiriyah's the " into "Nasiriyah".

### 3.4 Nominal Head Re-Tagging

One disadvantage of isolated mention translation is that heads cannot benefit from their contexts. For example, the head of "An eight-year-old Palestinian boy" is mistakenly translated into "boys", the head of "Palestinian security officials" is translated into "people". In order to address this problem, we take the following "head re-tagging" strategy: if the mention extent consists of "A of B" (ex. "Bureau of Religious") then we extract the last token of A as the head, otherwise consider the last token of the entire mention as the head.

### 3.5 Local Improvement

We apply lists of automatically extracted bilingual name pairs to re-translate some of the names; for some other names, we decompose the source name and translate the name components (for example, for company names, the name and corporate suffix are processed separately). Translations based on these lists take precedence over the MT output. The following lists of name pairs are used:

**LDC Person Name Lists.** Famous names (486212 names) and who's who international (36881 names).

**Name Pairs from ET Training Corpus.** Manually extracted from ET newswire training corpus, including 800 Person names, 448 GPE names, and 323 ORG names.

**Mention Pairs from ET Training Database.** Extracted and manually cleaned/corrected from ET training database, including 133 Peron names, 77 ORG names, 107 GPE names, 172 Person nominals, 43 ORG nominals and 26 GPE nominals.

These lists particularly helped to translate those non-famous names which share characters with famous names. For example, the MT system mistakenly translated "布哈里" into "Bush" because they share the same first character in Chinese, but by using name pairs it can be corrected into "Bohali".

In addition, we apply a tool[1] for converting Chinese character to pinyin in order to generalize the Chinese names, so we can 'cluster' the names into pinyins and then map them against lists. The conversion tool is based on a deterministic algorithm (i.e. no disambiguation for multiple pinyins) taken from the Unihan database [2] (unicode version 4.1.0, table version 1.1, Mar 29, 2005 release).

### 3.6 Source Language Coreference based Voting

After doing local improvement, we identify name coreference relations in the source language and compare the translations of coreferring mentions, applying a voting cache operating over a large set of machine-translated documents. The source language coreference information is thus encoded as a confidence metric and the high-confidence best translation can be propagated through the corpus, replacing lower-confidence translations.

In such a framework we address the noise produced by the MT system by using confidence estimation based on voting among translations of coreferring mentions. We build the following voted cache models in order to get the best *assignment* (translation candidate) for each entity:

**Inside-S-D-Cache**: For each name mention in one entity (inside a single document), record its unique translations and frequencies;

**Cross-S-D Cache**: Corpus-wide (across documents), for each name and its consistent variants, record its unique translations and their frequencies;

For each entry in these caches, we get the frequency of each unique *assignment*, and then use the following *margin* measurement to compute the confidence of the best assignment:

- *Margin = Frequency (Best Assignment) –Frequency (Second Best Assignment)*

A large margin indicates greater confidence in the assignment. Therefore, we pick up the most common translation with high *confidence*; and propagate it through the coreferred name mentions or the test corpus (across documents). This method particularly helps to correct the translations for GPE abbreviation names such as "叙 (Syria)", "拉美 (Latin America)" which otherwise will be missed or mistakenly translated into non-names. It also successfully translates more non-famous names because of the global propagation of correct assignments.

### 3.7 Averaged Perceptron Name Transliteration

We trained a sequence alignment model based on the averaged perceptron [Collins 2002] using a list of corresponding pinyin/English name token pairs, where the names are non-Chinese person names. For each training pair, the indicated English rendering constitutes our true target (t) and we use the current model to

[1] This pinyin conversion tool is publicly available at http://www.cs.nyu.edu/~yusuke/tools/pinyin.py

[2] http://www.unicode.org/Public/UNIDATA/Unihan.html

generate an alternate string (t'), updating the model in the event t' yields a higher score than t. This was repeated for 10 epochs. We used a model of order 3. Our model is described in greater detail in [Freitag and Khadivi 2007].

When testing the same approach with Arabic→English transliteration, we observed a 1-best accuracy of 0.552, a 5-best accuracy of 0.803, and a CER of 0.126. It is difficult to characterize the strength of these numbers relative to those reported in the literature. [Al-Onaizan and Knight 2002] reports a 1-best accuracy of 0.199 and a 20-best accuracy of 0.317 on a corpus of Arabic *person* names (but a 1-best/20-best of 0.634/0.852 on *English* names), using a "spelling-based" model, i.e., a model which has no access to phonetic information. However, the details of their experiment and model differ from ours in a number of respects.

We utilized the Chinese→English model to rank each Chinese source person name token identified by the Chinese EDR system against a list of English name tokens that includes Chinese and non-Chinese names. The English name tokens were extracted from a set of documents that conforms to the ACE-ET "blackout dates".

It is possible to use the highest-ranked English token as the translation of each Chinese source name token. However, we found that whereas the correct translation is commonly among the top-ranked English tokens, it is often not the very top one. For example, for the second token in "托尼.布莱尔" ("Tony Blair"), the model ranks English tokens as follows:

| | | | |
|---|---|---|---|
| bril | 27.36 | boulil | 22.15 |
| boer | 26.17 | butler | 21.77 |
| blair | 24.72 | breyer | 21.60 |
| blail | 24.64 | boolell | 19.80 |
| brill | 24.49 | burrell | 19.77 |

"Blair" is only third-best. Thus, we utilize the token ranking as input to a second stage of processing that utilizes multi-token name and document context, as described in the next section.

### 3.8 Cross-document Coreference with English Documents

If a name might be translated several ways, we would like to determine whether any of the candidate translations appears in English source documents [Al-Onaizan and Knight 2002] and, if so, choose the most plausible one. We consider four factors in determining the degree of fit:

- Edit distance (as described in Section 3.7)
- Multi-token names
- Name frequency in the English corpus
- Document context

Multi-token names provide a strong indication of the correct translation. If "Tonny Breyer" and "Tony Blair" are both feasible translations based on edit distance but only the latter appears in the English corpus, the latter is much more likely to be correct than the former. The computational complexity of our process for finding all multi-token matches is $O(N^2)$ in the number of matches to each individual component.

If multiple multi-token candidate translations appear in the English corpus, the more frequent ones are more likely to be correct than the less frequent ones (based on the assumption that the name occurrence statistics are similar in the source and English corpora).

Finally, if a candidate name appears in an English document published the same year (temporal proximity) as the source document, with the same topic (document classification) as the source document, or with co-occurring names shared by the source document, it is more likely to be the correct translation.

In the ACE-ET evaluation, we engineered a score combining edit distance, name frequency, and temporal proximity for multi-token names with candidate translations in the English corpus and selected the highest-ranked candidate based on this score. Utilizing additional types of document context would improve the selection accuracy and remains as future work.

Once a translation is selected for the longest form of a person name, we select the corresponding English sub-names for shorter named mentions of the same entity based on the source language coreference chain.

The number of "longest form of person name" translated by this method in the newswire and weblog portions of the evaluation data were 133 and 31, respectively. The number of shorter co-referring mentions translated based on this method in the newswire and weblog portions of the evaluation data were 153 and 18, respectively.

### 3.9 MT Name Transliteration

The other Chinese name transliteration module uses a system similar to [Al-Onaizan and Knight, 2002]. Their system can be regarded as a statistical machine translation (SMT) system which translates (transliterates) source language characters to target language characters in the absence of phonetic information.

We use the RWTH phrase-based statistical machine translation (SMT) system [Zens and Ney, 2004] to build a Chinese-to-English transliteration system. This system frames the transcription problem as follows. We are given a sequence of source language characters representing a name, which is to be translated into a sequence of target language characters. In comparison to the MT system described in Section 2.1, the transliteration system includes the following models: a character-level phrase translation model, a character-level lexicon model, an n-gram character sequence model, a character penalty and a phrase penalty. The first two models are used for both directions. We do not use any reordering model because the target character sequence is always monotone with respect to the source character sequence. We should note that each Chinese character is represented by its pinyin representation to the MT system.

To improve the transliteration results, we force the MT system to generate a sequence of tokens which are likely to be a name. To do this, we automatically extract and clean a large list of name entities from monolingual English corpora, then we select the best hypothesis for a Chinese name from an n-best list generated by the MT system according to the name entity list. We have also used this list in addition to the target side of the bilingual corpus to build a 6-gram character sequence model.

The results of this MT name transliteration are currently used as a "dynamic" electronic name translation dictionary. We first apply a cost threshold (3.6 in the current system) derived from the development set to select the high-confidence entries, and then if a name or its component does not conflict with any of the name pairs got from ET training data, we consider it as a "valid" entry and then apply the translation of this name or its component through the whole test corpus.

### 3.10 Mention Extent Smoothing

The definitions of the extents for apposition name mentions are different between Chinese EDR and English EDR, for example, in the name mention "Prime Minister Atef Obeid", Chinese EDR defines the extent as "Atef Obeid" but English EDR as "Prime Minister Atef Obeid". So after we get translations from the above pipelines, we extend the boundaries of name extents by merging with apposition nominals immediately before them, if there are any.

In addition, another disadvantage of isolated mention translation is that it tends to miss "the" at the beginning of some mention extents (ex. "(the) British company"). We address this by searching for the English mention in the MT output for the whole document; if "the" precedes the candidate mention string then we add it in the final output.

## 4 Pipeline Combination

Above, we have described two methods for generating English entities of a Chinese document – Pipeline 1: target language oriented ET; and Pipeline 2: source language oriented ET and its three "sub-pipelines": Pipeline 2.1, Pipeline 2.2 and Pipeline 2.3 as shown in Figure 1.

In the ET-07 evaluation we use Pipeline 2.2 to override the results of Pipeline 2.1, and use high-confidence translations produced by Pipeline 2.3 to override those mentions that don't get results from lists and voting in Pipeline 2.1 or Pipeline 2.2. In the future we shall try different priorities in combination; and shall try to learn priorities based on the confidence values produced by these pipelines; or to train a statistical learner to rank/combine them.

Some entities overlapped in Pipeline 1 and Pipeline 2 results but with different entity types. We apply the following rule optimized from the development set: if Pipeline 1 identifies the entity as a location, weapon or vehicle then remove it from the results of Pipeline2; else if Pipeline 1 identifies it as a facility entity then remove it from Pipeline 1.

In the diagnostic ET evaluation we didn't use Pipeline1, instead we used Pipeline 2 to produce all types of entities because of the perfect entities given in Chinese.

## 5 Feedback from ET to Improve Chinese EDR

We also use the 'cross-lingual feedback' from individual mention translations to improve Chinese EDR. We test properties of the translated mentions, as well as comparing the translations of coreferring mentions. The intuition is that if a correct Chinese sequence of tokens is correctly identified as a name, it is more likely to be translated into a name (generally, a sequence of capitalized tokens) in English. So if we have competing name tagging hypotheses, we use the high-confidence translations to select the correct analysis. For example we can correct the Chinese name boundary of "三菱新(*Mitsubishi new*)" into "三菱" because of the lower-case translation for "new". We build:

**Cross-D-S cache**: Corpus-wide, for each consistent English translation, record its corresponding names in Chinese and their frequencies.

This cache, together with the other caches described in section 3.6, incorporate simple filters based on English properties to exclude translations which are not likely to be names. We exclude empty translations, translations which are single un-capitalized tokens, and, for person names, translations with any

un-capitalized tokens. In addition, in counting translations in the cache, we group together consistent translations. This includes combining person name translations if one is a subsequence of the tokens in the other. The goal of these simple heuristics is to take advantage of the general properties of English in order to increase the likelihood that the most frequent entry in the cache is indeed the best translation.

We combine the language-specific information in Chinese entities, and its entry in the cross-lingual caches to detect potential extraction errors and take corresponding corrective measures. We construct some plausible inference rules that improve Chinese mention tagging, Chinese coreference and entity translation. These rules are applied repeatedly until there are no further changes; improved translation in one iteration can lead to improved Chinese entity extraction in a subsequent iteration.

We didn't apply this method in the diagnostic ET evaluation because the source language entities are supposed to be perfect.

## Acknowledgement

We would like to thank Yusuke Shinyama for providing the pinyin conversion tool and the Chinese coding conversion tool, Shasha Liao for correcting training name pairs. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. 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 Defense Advanced Research Projects Agency (DARPA).

## References

[Al-Onaizan and Knight 2002] Yaser Al-Onaizan and Kevin Knight. **Translating Named Entities Using Monolingual and Bilingual Resources**. ACL 2002.

[Bikel et al. 1997] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. **Nymble: a high-performance Learning Name-finder.** ANLP 1997.

[Collins 2002] Michael Collins. **Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms**. EMNLP-2002.

[Freitag and Khadivi 2007] Dayne Freitag and Shahram Khadivi. **A Sequence Alignment Model Based on the Averaged Perceptron**. Submitted to ACL 2007.

[Grishman et al. 2005] Ralph Grishman, David Westbrook and Adam Meyers. **NYU's English ACE 2005 System Description**. ACE 2005 PI Workshop.

[Ji and Grishman 2005] Heng Ji and Ralph Grishman. **NYU's Chinese ACE 2005 EDR System Description**. ACE 2005 PI Workshop.

[Och 2003] F.J. Och. **Minimum Error Rate Training in Statistical Machine Translation.** ACL 2003.

[Sudo et al. 2004] Kiyoshi Sudo, Satoshi Sekine and Ralph Grishman. **Cross-lingual Information Extraction System Evaluation**. COLING 2004.

[Zens and Ney 2004] Richard Zens and Hermann Ney. **Improvements in Phrase-Based Statistical Machine Translation.** HLT/NAACL 2004.