ICE-TEA: In-Context Expansion and Translation of English Abbreviations

Waleed Ammar, Kareem Darwish, Ali El Kahki, and Khaled Hafez[1]

Cairo Microsoft Innovation Center, Microsoft, 306 Chorniche El-Nile, Maadi, Cairo, Egypt
{i-waamma,kareemd,t-aleka}@microsoft.com,
hafez.khaled@gmail.com

Abstract. The wide use of abbreviations in modern texts poses interesting challenges and opportunities in the field of NLP. In addition to their dynamic nature, abbreviations are highly polysemous with respect to regular words. Technologies that exhibit some level of language understanding may be adversely impacted by the presence of abbreviations. This paper addresses two related problems: (1) expansion of abbreviations given a context, and (2) translation of sentences with abbreviations. First, an efficient retrieval-based method for English abbreviation expansion is presented. Then, a hybrid system is used to pick among simple abbreviation-translation methods. The hybrid system achieves an improvement of 1.48 BLEU points over the baseline MT system, using sentences that contain abbreviations as a test set.

Keywords: statistical machine translation, word sense disambiguation, abbreviations.

1 Introduction

Abbreviations are widely used in modern texts of several languages, especially English. In a recent dump of English Wikipedia,[2] articles contain an average of 9.7 abbreviations per article, and more than 63% of the articles contain at least one abbreviation. At sentence level, over 27% of sentences, from news articles, were found to contain abbreviations. The ubiquitous use of abbreviations is worth some attention. Abbreviations can be acronyms, such as NASA, which are pronounced as words, or initialisms, such BBC, which are pronounced as a sequence of letters.

Often abbreviations have multiple common expansions, only one of which is valid for a particular context. For example, Wikipedia lists 17 and 15 valid expansions for IRA and IRS respectively. However, in the sentence: *"The bank reported to the IRS all withheld taxes for IRA accounts."* IRA conclusively refers to *"individual retirement account"* and IRS refers to *"internal revenue service"*. Zahariev (2004) states that 47.97% of abbreviations have multiple expansions (at WWWAAS[3])

[1] Author was an intern at Microsoft and is currently working at the IBM Technology Development Center in Cairo.

[2] http://dumps.wikimedia.org/enwiki/20100312/

[3] World-Wide Web Acronym and Abbreviation Server
http://acronyms.silmaril.ie/

A. Gelbukh (Ed.): CICLing 2011, Part II, LNCS 6609, pp. 41–54, 2011.

compared to 18.28% of terms with multiple senses (in WordNet), suggesting that abbreviations are highly polysemous with respect to regular words. Table 1 lists some popular abbreviations with multiple expansions.

Table 1. Some popular polysemous abbreviations

Abb.	Expansion
TF	Term Frequency
TF	Task Force
IDF	Israel Defense Forces
IDF	Inverse Document Frequency
IDF	Intel Developers Forum
CIA	Central Intelligence Agency
CIA	Certified Internal Auditor
IRA	Irish Republican Army
IRA	Individual Retirement Account
AP	Advanced Placement
AP	Associated Press
AP	Access Point
ATM	Asynchronous Transfer Mode
ATM	Automated Teller Machine
ATM	Air Traffic Management

Abbreviations pose interesting challenges and opportunities in Statistical Machine Translation (SMT) systems such as (Koehn et al., 2003; Quirk et al. 2005; Galley et al., 2006; Chiang, 2007). Some of the challenges include:

1. The proper abbreviation translation may not exist in parallel data that was used for training. Given the dynamic aspect of abbreviations, where tens of new abbreviations appearing every day (Molloy, 1997), parallel text used for training may be limited or out-of-date. Typically, available parallel text hardly covers one or two (if any) common translations of an abbreviation, overlooking less common translations.
2. Many abbreviations are polysemous. Even in cases when multiple translations are observed in parallel training text, sufficient context is often not available to enable a language model to promote the proper translation.

Intuitively, an SMT system may have a better chance at translating an expanded abbreviation than the abbreviation itself. If an abbreviation can be properly expanded prior to translation, the ambiguity is removed (availing problem 2), and the MT system may be able to produce a reasonable translation even if it does not exist in training data (availing problem 1).

The contributions of this paper are:

1. an efficient Information Retrieval (IR) based technique for abbreviation expansion,
2. the use of abbreviation expansion to enhance translation of sentences that contain abbreviations, and
3. a hybrid system that picks from among four different abbreviation translation methods.

In this work, abbreviation expansion is treated as a retrieval problem using a probabilistic retrieval model to compute the similarity between observed context and each of existing contexts of expansions that share the same abbreviation. As for abbreviation translation, the hybrid system picks from: direct in-context abbreviation translation, in-context and out-of-context translation of expansion, and transliteration. The paper demonstrates the effectiveness of the proposed methods on English to Arabic MT. Unlike English, abbreviations are rare in Arabic.

Abbreviation expansion is a special case of Word Sense Disambiguation (WSD). However, abbreviations have characteristics that necessitate handling them differently. Unlike normal words, abbreviations have well defined senses. Also, it is relatively easy to get training documents that contain abbreviations along with their expansions. Most research on WSD addresses these two aspects of disambiguation (i.e. definition of word senses and sense-annotated corpora), which is not a major concern for disambiguation of abbreviations. In addition, given their dynamic nature, many abbreviations have low chance to appear in parallel data compared to normal words. Consequently, special approaches to disambiguate and then translate abbreviations are needed.

The remainder of the paper is organized as follows: Sections 2 and 3 explain the proposed approaches for abbreviation expansion and abbreviation translation respectively; Section 4 describes the experimental setup and reports on results; Section 5 provides related work in the literature; Section 6 concludes the paper and proposes future work.

2 Abbreviation Expansion

2.1 Problem Statement

Given text T which contains an ambiguous abbreviation α and given a set of possible expansions $E = \{e_1, e_2, \ldots e_n\}$, abbreviation expansion is defined as the problem of selecting the proper expansion $e_k \in E$ of α given T.

2.2 Retrieval-Based Solution

The proposed approach is based on the assumption that contextual text T relates to documents which contain the correct expansion e_k more than documents which contain other expansions $e_{i \neq k}$. For each abbreviation-expansion pair found in a document, the tuple {abbreviation, expansion, context} is constructed. Context refers to a set of sentences that contain the expansion for the abbreviation. The tuples are indexed offline using an IR engine. At query time, the index is queried using

significant terms in text T as keywords, restricting results to those where abbreviation = α. The highest ranking expansion is assumed to be the proper expansion e_k.

Introducing possible expansions methods is beyond the scope of this paper; interested readers can refer to (Yeates, 1999; Hiroko and Takagi, 2005; Larkey et al., 2000; Xu and Huang, 2006; Zahariev, 2004). In addition, several resources on the web maintain up-to-date abbreviation definitions and serve them for free (e.g. The Internet Acronym Server[4], Acronym Finder[5] and Abbreviations[6]).

Given a database of abbreviations and their possible expansions, it is straightforward to obtain training documents which contain a particular abbreviation expansion. Web search engines can be used for this purpose by specifying the abbreviation expansion as a phrase in addition to the abbreviation itself. However, since the authors did not have access to any database of abbreviation expansions, a method similar to that of Larkey et al. (2000) was used to identify abbreviations and their expansions in Wikipedia articles, creating a database of abbreviations and expansions (more details in section 4). The method relied on using heuristics to automatically identify abbreviations and their expansions in a large corpus. The corpus used herein was the English Wikipedia pages.

One of the advantages of using an IR engine is that, unlike binary discriminative classifiers, features (i.e. words in all contexts) assume consistent weights across classes (i.e expansions). Unlike most related work (e.g. Zahariev, 2004; Gaudan et al., 2005) where a classifier is built for each expansion requiring multiple classifiers to be used for each abbreviation, IR engine can ascertain the best expansion by quering one index.

For this work, retrieval was performed using Indri, a commonly used open source IR engine, which is based on inference networks and allows for fielded search (Metzler and Croft, 2004).

2.3 Learning Unseen Expansions

The proposed solution for abbreviation expansion cannot resolve abbreviations not defined in training documents. In order to keep the system up-to-date and complement shortages that may exist in training corpus, acquiring new tuples of abbreviation-expansion-context has to be an ongoing process. This is achieved by mining input text T to identify abbreviation definitions that may exist, in parallel to the normal processing described in the previous subsection (2.2). Texts which contain such tuples are incrementally indexed, and added to the training corpus for later use.

3 Abbreviation Translation

This section discusses several methods to translate a sentence S that contains an ambiguous abbreviation α. Given that different methods have advantages and disadvantages, a hybrid system that utilizes language modeling is used to pick from among the output of all methods.

[4] http://acronyms.silmaril.ie/
[5] http://www.acronymfinder.com/
[6] http://www.abbreviations.com/

3.1 Leave and Translate (LT)

This is the baseline. In this method, no special treatment is given to abbreviations. A syntactically informed phrasal SMT system, similar to that of Menezes and Quirk (2008) and Quirk et al. (2005) was used. This method performs well only with popular and unambiguous abbreviations (e.g. UNESCO, WWW), but it suffers from the problems mentioned in the introduction.

3.2 Expand and Translate in-Context (ETC)

In this method, abbreviations are expanded prior to translation. The rationale behind this method is that MT systems may have a better chance of translating an abbreviation expansion than translating the abbreviation itself. Usually, abbreviation expansions have reduced lexical ambiguity and improved lexical coverage as the constituents of an expansion are more likely to have relevant entries in the phrase-table compared to abbreviations. Also, expansion of abbreviations is informed by more context than language models which may only account for small windows of word n-grams. The proposed method works as follows:

1. Find the most likely expansion e_k of the abbreviation α given its context.
2. Replace α in the sentence S with e_k, producing modified sentence S`.
3. Translate the modified sentence S` using baseline MT system.

For example, consider the following two sentences:

S_1: ***ATM*** *is a networking protocol.*
S_2: *There's a nearby* ***ATM*** *in case you need to withdraw cash.*

Using the LT method (as in subsection 3.1) to translate the English sentences to Arabic leads to identical translations of *ATM* for both sentences:

LT(S_1): بروتوكول شبكة اتصال. **جهاز الصراف الآلي**
LT(S_2): يوجد **جهاز الصراف الآلي** قريبة في حال أردت سحب المال.

In contrast, ETC first transforms the English source sentences to:

$S`_1$: ***Asynchronous transfer mode*** *is a networking protocol.*
$S`_2$: *There's a nearby* ***automatic teller machine*** *in case you need to withdraw money.*

Then, ETC translates the modified sentences, producing a much better translation for *ATM* in the first sentence:

ETC(S_1): بروتوكول شبكة اتصال. **وضع النقل غير المتزامن**
ETC(S_2): يوجد **جهاز الصراف الآلي** قريبة في حال أردت سحب المال.

A drawback of this method is that the MT decoder may inappropriately breakup phrases to match against the phrase-table. For example, the decoder may decide to translate *"nearby automatic"* and *"machine in case"* as phrases.

3.3 Expand and Translate Out-of-Context (ETOC)

To avoid the drawback described in 3.2, this method gains partial control over the segmentation of modified source sentences by translating the expansion in isolation, and then replacing the abbreviation in the source sentence prior to translation by the MT engine, as follows:

1. Find the most likely expansion e_k for the abbreviation α given its context *(identical to ETC's step 1)*.
2. Translate the most likely expansion e_k in isolation to target language phrase A.
3. Replace the abbreviation α in the source sentence with an OOV word, producing modified sentence S`.
4. Translate S`, *producing* T.
5. Replace the OOV word in T by A.

Building on the *ATM* example, the isolated translations of the expansions (step 2) produce:

A_1: وضع النقل غير المتزامن
A_2: جهاز الصراف الالي

Replacing the abbreviation α with translation A (step 3) produces:

$S`_1$: **OOV** *is a networking protocol.*
$S`_2$: *There's a nearby* **OOV** *in case you need to withdraw money.*

Translating the modified sentence S` (step 4) produces:

T_1: بروتوكول شبكة اتصال. **OOV**
T_2: يوجد **OOV** قريبة في حال أردت سحب المال.

Replacing the **OOV** word with the expansion translation **A** (step 5) produces:

$ETOC(S_1)$: بروتوكول شبكة اتصال. **وضع النقل غير المتزامن**
$ETOC(S_2)$: يوجد **جهاز الصراف الآلي** قريبة في حال أردت سحب المال.

One caveat that limits the usefulness of this method is that the introduction of out-of-vocab words confuses the translation model. In order to reduce dependence on any particular decoder and to enhance reproducibility of this work, authors preferred not to introduce changes to the decoder to address this issue.

3.4 Transliteration (TT)

Some abbreviations are sufficiently popular that people use the abbreviated form in several languages. For example, the most common Arabic translation of NASA and BBC are ناسا and بي بي سي respectively. In such cases, transliteration can be the preferred translation method. When a popular abbreviation is an acronym (e.g. NASA, AIDS), the phonetically equivalent word in Arabic is a borrowed word (NASA→ناسا, AIDS→إيدز). When a popular abbreviation is an initialism[7] (e.g. BBC, AFP), a letter-by-letter transliteration is usually the most common translation (e.g. AFP→أ ف ب, BBC→ بي بي سي).

[7] Despite the difference between acronyms and initialisms, people often refer to both as acronyms.

In order to find the most common Arabic transliteration, Table 2 was used to produce possible transliterations of initialisms and acronyms. In short, English letters were replaced by their corresponding phonetic equivalents and then a language model (trained exclusively on target-language named-entities from publicly available ACE[8] and Bin-Ajeeba[9] corpora) was consulted to select the most likely transliteration. Phonetic transliteration of acronyms is left for future work.

Table 2. Arabic mappings of English letters

English letter(s)	Mappings for acronyms	Mappings for initalisms	English letter(s)	Mappings for acronyms	Mappings for initialisms
A	أا	ايه	N	ن	إن
B	ب	بي	O	وأ	أو
C	س|ك	سي	P	ب	بي
D	د|ض	دي	Q	ك|ق	كيو
E	إ|ي	إي	R	ر	آر
F	ف	إف	S	س|ز	إس
G	ج	جي	T	ت	تي
H	ه	إتش	U	و|يو	يو
I	ي	أي	V	ف	في
J	ج	جيه	W	و	دبليو
K	ك	كي	X	س|ك	إكس
L	ل	إل	Y	ي	واي
M	م	إم	Z	ز|س	زد

3.5 Hybrid (HYB)

None of the aforementioned methods is expected to consistently yield the best translation results. For example, if an abbreviation α appears (with the sense e_k) a sufficient number of times in the parallel training data and the general language model can properly pick the proper translation, then the LT method is likely to produce a correct translation. If the abbreviation is not present in the parallel training data, but its constituents do, methods ETC and ETOC are expected to produce better translations. If the abbreviation is used at the target language as a borrowed word, TT would be the method of choice.

The hybrid method translates the sentence using the four methods (LT, ETC, ETOC and TT) and selects the most fluent translation (as estimated by target language model probability).

[8] http://projects.ldc.upenn.edu/ace/
[9] http://www1.ccls.columbia.edu/~ybenajiba/

4 Experiments

4.1 Abbreviation Expansion

In this work, abbreviation expansion was examined for the English language. English Wikipedia articles were scanned for abbreviation-expansion pairs. An abbreviation-expansion pair was extracted when an expansion was followed by an abbreviation between brackets, where letters in the abbreviation matched sequences of initial and middle letters of words in the expanded form. Frequency of an abbreviation-expansion pair has to surpass a threshold (3 was used as the threshold) to qualify as a valid pair. As a by-product of this process, example documents containing the abbreviations and their expansions were automatically obtained. In all, the constructed collection contained unique 10,664 abbreviations with 16,415 unique expansions, extracted from roughly 2.9 million Wikipedia articles. The number of expansions per abbreviation was 1.54 on average with a variance of 1.66.

Context documents were indexed using Indri fielded indexing, with the fields being the abbreviation, the expansion, and full text of the document.

The test set of abbreviations, expansions and contexts contained 500 English Wikipedia articles, randomly extracted and manually revised. Mentions of the abbreviation expansions were removed from the context. For testing, the context query for an abbreviation was taken as the 10 words preceding, 10 words trailing the abbreviation (excluding stopwords). If an abbreviation appeared more than once in the article, context words were aggregated for all occurrences. The 50 unique words with highest term frequency in the article were selected as the context query for the abbreviation. The query was submitted to Indri constraining the results while restricting the abbreviation field to the abbreviation at hand. The expansion corresponding to the highest ranking result was chosen as the proper expansion. Accuracy was used as the figure of merit to evaluate abbreviation expansion. When calculating accuracy, the system received a 1 if the top returned expansion was correct and 0 otherwise.[10]

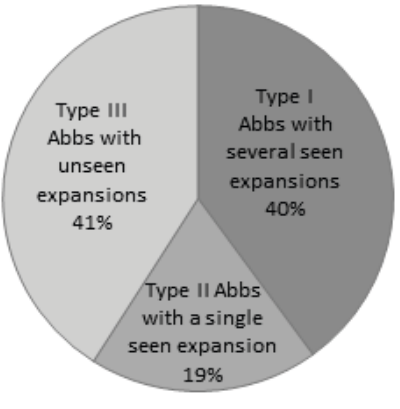

Fig. 1. Breakdown of test set types

[10] This is referred to as precision@1 in IR literature.

Figure 1 provides a breakdown of abbreviations in the test set as follows: Type I: 202 (out of 500) polysemous abbreviations had an average of 3.7 possible expansions each; Type II: 94 abbreviations had only 1 expansion (so the proposed approach did not have a choice but to select the correct expansion); and Type III: 204 abbreviation-expansion pairs were not previously observed (hence the proposed technique had no chance of finding them). Table 3 presents the results obtained for abbreviation expansion. The results reported here include accuracy when all test data is used (types I, II, and III), when excluding test items for which the correct expansion was never seen in training data (types I, II), and when excluding abbreviations which have a single possible expansion as well (type I). The baseline reported here is a process that selects an expansion randomly assuming uniform distribution for the possible expansions. Unfortunately, the authors did not have access to datasets reported on in the literature for comparative evaluation.

Table 3. Accuracy of abbreviation expansion

Test set	Baseline	IR
All (500 test tuples)	35%	53%
Types I & II (296 tuples)	44%	90%
Type I only (202 tuples)	27%	86%

When considering the abbreviations for which no expansions were seen in training, the proposed approach achieved 53% accuracy. Overcoming such low performance would require increasing the training data by learning new abbreviation-expansion pairs as explained in section 2. When excluding expansions that were missing from the training data, the system yielded 90% accuracy, which would be typical if the training set was to be expanded. It is noteworthy that when examining the mistakes that were made by the system, they included several examples such as "Singapore Art Museum" where the system's guess was "Seattle Art Museum"; such examples are probably harder to disambiguate than others. When abbreviations with a single expansion were excluded, the accuracy was 86%.

4.2 Abbreviation Translation

For abbreviation translation, the aforementioned translation methods were tested individually as well as the hybrid method with different combinations. All experiments were performed using a syntactically informed statistical MT system similar to (Menezes and Quirk, 2008). Performance was evaluated using BLEU score, as measured by NIST's mteval (v. 13). Parallel training data was comprised of LDC news corpora[11], UN data, as well as automatically mined parallel sentences from the web (Quirk et al., 2007). A total of 11M sentences were used in training, with an average English and Arabic sentence lengths of 23 and 27 tokens respectively.

[11] LDC2004T18, LDC2004T17, LDC2007T24, LDC2008T02, LDC2008T09 and LDC2009T09.

500K parallel sentences were held out for testing and tuning purposes. An automatic process extracted 756 sentence pairs from the held out parallel data such that a unique abbreviation exists in each source sentence. Out of those, 500 sentence pairs were used as a test set, the rest were added to the development set. The test set had average English and Arabic sentence lengths of 30 and 29 respectively. Note that the unique abbreviations condition imposed a limit on the size of the test set. Further, BLEU scores reported here are lower than typical Arabic-to-English scores due to the lack of multiple reference translations for the English-to-Arabic test set.

The test set used for abbreviation translation was different than the one used for abbreviation expansion (Section 4.1). In abbreviation expansion, each test sample must identify the proper expansion of the abbreviation, which is not available for abbreviations in the parallel test set. On the other hand, abbreviation translation requires each test sample to contain the proper translation of the sentence, which is not available for Wikipedia articles used to test abbreviation expansion.

SRILM toolkit was used to build an Arabic trigram language model with Good-Turing smoothing for the hybrid method. The language model was constructed using texts in Arabic Wikipedia, Arabic Giga Word collection[12], and the Arabic portion of training data.

Table 4 lists the results of using the aforementioned methods for translation of sentences that contain abbreviations. While individual methods, namely ETC and ETOC, showed a small improvement over the baseline (LT), the hybrid system effectively combined translations from all four methods to achieve a significant improvement of 1.48 BLEU points. Using the hybrid method to pick among different methods consistently gave better results than individual methods, suggesting that target language model was effective in combine several abbreviation translation methods.

Table 4. BLEU score for abbreviation expansion translation using different methods

Method/Combination	BLEU
Baseline (LT)	16.60
ETC	17.01
ETOC	16.98
TT	14.65
Hybrid (LT, ETC)	17.35
Hybrid (LT, ETOC)	17.70
Hybrid (LT, ETC, TT)	17.27
Hybrid (LT, ETOC, TT)	17.68
Hybrid (LT, ETC, ETOC)	18.04
Hybrid (LT, ETC, ETOC, TT)	18.08

[12] LDC2003T12.

However, individual methods contributed differently to this improvement. Implementing the hybrid method using different combinations helps analyze the contribution of each method. Combinations (LT, ETC) and (LT, ETOC) gave improvements of 0.75 and 1.10 BLEU points, respectively. This confirms the assumption that expanding abbreviations before translation is beneficial.

Adding transliteration (TT) to any combination seemed to either degrade or yield (almost) the same BLEU score. This is probably attributed to the type of abbreviations for which TT was designed. It was expected to produce meaningful results for popular abbreviations. Nevertheless, such popular abbreviations are also expected to appear frequently in parallel training data. Consequently, baseline MT system would be sufficient to find the proper translation of such abbreviations in the phrase table, refuting the need to use TT.

One factor that limited the gain of abbreviation expansion methods was the prevalence of sentences in the test set where abbreviations were not fully translated. For example, the reference translation for the sentence containing the abbreviation KAC (Kuwait Airways Corporation) only referred to KAC as "the corporation" (الشركة). While this translation is valid when the full name is written earlier in the text, the translation is not complete in isolation. One way to avoid this problem is to perhaps use multiple reference translations, or to manually create the references in isolation of the full context of the documents.

5 Related Work

The problem of abbreviation expansion can be viewed as a special case of word sense disambiguation (WSD) (Zahariev, 2004). Over the past sixty years, sophisticated approaches were developed to address WSD. Interested readers are referred to a recent comprehensive survey on WSD by Navigli (2009). Although polysemy (i.e. lexical ambiguity) in abbreviations is often greater than polysemy in regular words (Zahariev, 2004), the representation of word senses in abbreviations is less of a problem than in regular words. For instance, most people would distinguish [gold: noun] and [gold: adjective] as different senses, but some people will go further and argue that [gold: of the color of gold] and [gold: consisting of gold] should be two distinct senses as well. Fortunately, this problem almost does not exist for abbreviations, making it more feasible to find a satisfactory solution to the problem given available resources.

Several supervised and semi-supervised learning approaches were used to solve the abbreviation expansion problem. In general text, Zahariev (2004) used a support vector machine (SVM) classifier with a linear kernel. A model is trained for each abbreviation, with distinct expansions representing different classes. Terms occuring in the same document as the abbreviation were used as features. Training data were obtained by searching the web for PDF documents containing both an abbreviation and any of its expansions. Though effective, building SVM models for each expansion of every abbreviation was computationally intensive. SVM attempted to assign different weights to different features and these weights were different from one model to the next.

Solving this problem for the medical domain captured the interest of many researchers due to the widespread use of abbreviations in the biomedical domain. Pakhomov et al. (2002) approached the problem using maximum entropy classification, Gaudan et al. (2005) used an SVM classifier, and Stevenson et al. (2009) used a vector space model.

Roche and Prince (2008) ranked the expansions of a given abbreviation by calculating the cubic mutual information function and Dice's coefficient based on the number of web pages. Given an abbreviation, contextual text, and the set of possible expansions, their idea was to find the number of web pages containing the expansion and keywords from the context, then to divide this value by the number of pages containing individual key/expansion words. The evaluation was done for French as well as medical abbreviations.

Some research efforts targeted translation of abbreviations. Callison-Burch et al. (2006) looked at the broader problem of using paraphrases to improve lexical coverage in MT. Along the same line, Li and Yarowsky (2008a; 2008b) used an unsupervised technique to address translation of Chinese abbreviations. English named entities (NEs) were extracted from a monolingual corpus, and translated using a baseline MT system into Chinese. Then, Chinese abbreviation-expansion pairs were extracted from monolingual Chinese text, and matched with their English NE translations using the Chinese automatic translation obtained before as a bridge. Then, Chinese abbreviations and their corresponding English NE translations were added to the phrase table of the baseline MT system. While this approach effectively solved the first problem mentioned in the introduction (i.e. the proper phrase pair does not exist in the phrase table), it does not solve the second problem (i.e. the high polysemy of abbreviations) because the decoder was still responsible for disambiguating between multiple expansions (i.e. translations) of a polysemous abbreviation using the target language model. On the other hand, the proposed approach at hand addresses English abbreviations, solves both identified problems, and experiments with several methods of translating an abbreviation.

Some researchers (Chan et al., 2007; Carpuat and Wu, 2007) also studied the integration of WSD in SMT by introducing elaborate modifications to the decoders. In this work, although abbreviation expansion was a special case of WSD, a design decision was taken to simplify the integration by using the decoder as a black box, making it much easier to implement, replicate and scale to different SMT systems.

6 Conclusion and Future Work

A retrieval-based algorithm for abbreviation expansion was presented. Using a retrieval engine for abbreviation expansion availed the need to build separate classification models for different abbreviation. The described algorithm was both efficient and effective, yielding an accuracy of 90%.

Regarding translation, expanding abbreviations before translation was a simple but useful modification. A hybrid system that utilized a variety of abbreviation translation methods was presented. While individual methods showed small improvements, combining several methods achieved significant improvement of 1.48 BLEU points.

This work can be extended in three directions. One direction is to generalize the proposed IR-based disambiguation technique for words rather than abbreviations. The main difficulty here lies in the definition of word senses and developing sense-annotated corpora. The second direction is to enhance the proposed abbreviation translation approach. In particular, a proper way to condense translated abbreviation expansions is needed. For example, a professional translator would translate English "UN" into French "ONU", while the proposed approach would translate it to French "Organisation des Nations Unies". Also, using acronym phonetic transliteration may make the TT method more effective. The third direction is to make use of abbreviation expansion in other IR/NLP tasks that exhibit some sort of language understanding (e.g. query expansion and question answering).

Acknowledgements

Authors would like to thank Arul Menezes and Mei-Yuh Hwang for their scientific and technical support, Hany Hassan and Khaled Ammar for their numerous valuable comments, and Amany Shehata for performing human evaluations.

References

Agirre, E., Martinez, D.: Smoothing and word sense disambiguation. In: Vicedo, J.L., Martínez-Barco, P., Muńoz, R., Saiz Noeda, M. (eds.) EsTAL 2004. LNCS (LNAI), vol. 3230, pp. 360–371. Springer, Heidelberg (2004)

Callison-Burch, C., Koehn, P., Osborne, M.: Improved Statistical Machine Translation Using Paraphrases. In: NAACL 2006 (2006)

Carpuat, M., Wu, D.: Improving Statistical Machine Translation using Word Sense Disambiguation. In: Proceedings of EMNLP 2007, pp. 61–72 (2007)

Chan, Y., Ng, H., Chiang, D.: Word Sense Disambiguation Improves Statistical Machine Translation. In: Proceedings of ACL 2007, pp. 33–40 (2007)

Chiang, D.: Hierarchical Phrase-Based Translation. Computational Linguistics 33(2), 201–228 (2007)

Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., Thayer, I.: Scalable Inference and Training of Context-Rich Syntactic Translation Models. In: Proceedings of COLING/ACL 2006, pp. 961–968 (2006)

Gaudan, S., Kirsch, H., Rebholz-Schuhmann, D.: Resolving abbreviations to their senses in Medline. Bioinformatics 21(18), 3658–3664 (2005)

Hiroko, A., Takagi, T.: ALICE: An Algorithm to Extract Abbreviations from MEDLINE. Journal of the American Medical Informatics Association, 576–586 (2005)

Koehn, P., Och, F., Marcu, D.: Statistical phrase-based translation. In: Proceedings of NAACL 2003, pp. 48–54 (2003)

Larkey, L., Ogilvie, P., Price, A., Tamilio, B.: Acrophile: an automated acronym extractor and server. In: Intl. Conf. on Digital Libraries archive, 5th ACM Conf. on Digital libraries, pp. 205-214 (2000)

Li, Z., Yarowsky, D.: Unsupervised Translation Induction for Chinese Abbreviations using Monolingual Corpora. In: ACL 2008 (2008a)

Li, Z., Yarowsky, D.: Mining and modeling relations between formal and informal Chinese phrases from web corpora (2008b)

Menezes, A., Quirk, C.: Syntactic Models for Structural Word Insertion and Deletion. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 735–744, Honolulu (October 2008)

Metzler, D., Croft, W.B.: Combining the Language Model and Inference Network Approaches to Retrieval. Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval 40(5), 735–750 (2004)

Molloy, M.: Acronym Finder (1997), from http://www.acronymfinder.com/ (retrieved February 8, 2010)

Navigli, R.: Word Sense Disambiguation: a Survey. ACM Computing Surveys 41(2) (2009)

Pakhomov, S.: Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts. In: ACL 2002 (2002)

Quirk, C., Menezes, A., Cherry, C.: 2005. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In: ACL 2005 (2005)

Quirk, C., Udupa, R., Menezes, A.: Generative Models of Noisy Translations with Applications to Parallel Fragment Extraction. In: European Assoc. for MT (2007)

Roche, M., Prince, V.: AcroDef: A quality measure for discriminating expansions of ambiguous acronyms. In: Kokinov, B., Richardson, D.C., Roth-Berghofer, T.R., Vieu, L. (eds.) CONTEXT 2007. LNCS (LNAI), vol. 4635, pp. 411–424. Springer, Heidelberg (2007)

Roche, M., Prince, V.: Managing the Acronym/Expansion Identification Process for Text-Mining Applications. Int. Journal of Software and Informatics 2(2), 163–179 (2008)

Stevenson, M., Guo, Y., Al Amri, A., Gaizauskas, R.: Disambiguation of Biomedical Abbreviations. In: BioNLP Workshop, HLT 2009 (2009)

Xu, J., Huang, Y.: Using SVM to Extract Acronyms from Text. In: Soft Computing - A Fusion of Foundations, Methodologies and Applications, pp. 369–373 (2006)

Yeates, S.: Automatic Extraction of Acronyms from Text. In: New Zealand Computer Science Research Students Conference 1999, pp. 117–124 (1999)

Zahariev, M.: Automatic Sense Disambiguation for Acronyms. In: SIGIR 2004, pp. 586–587 (2004)