Cool-to-Hot Pivoting for Improved Machine Translation

Rob Voigt – CS224N, Fall 2012

Final Project

Note:

This project presents work in progress with Daniel Cer and Dan Jurafsky; however, all associated code, annotation, linguistic and error analysis, and writing thus far has been done by me alone. Their contribution to this point has been through weekly progress meetings and group brainstorming. Also, since we are planning to submit a later version of this research to ACL in February, I ask that you not post this project report on the CS224N website. Many thanks.

1. Introduction

Tsao (1977) argues that languages like Chinese can be seen as distinct from languages like English under the terms of what he describes as a "discourse-oriented vs. sentence-oriented" parameter. Other scholars encapsulate a similar distinction by means of a proposed "hot-cool" parameter, wherein "hot" languages such as English more explicitly express certain anaphoric elements whereas "cool" languages such as Chinese require greater inferational or contextual "work" on the part of a hearer to understand an utterance (Ross 1982, Huang 1984). "Sentence-oriented" or "cool" languages tend to present unique difficulties for natural language processing systems, which commonly operate, at least at some level, on the basisa of n-gram or sentence-level models. Therefore, this paper presents work in progress on a novel technique for explicitly integrating pragmatic linguistic knowledge into NLP systems, and particularly machine translation systems, for "cool" languages.

Chinese, Japanese, and Korean are classic examples of "pro-drop languages"; that is, pronouns are often omitted in both spoken and written language when they can be inferred pragmatically. Since almost all modern machine translation systems operate on a sentence-by-sentence basis, a null argument cannot be recovered and is necessarily translated randomly or not at all. Moreover, since the most common automatic metrics that MT researchers are hillclimbing, such as BLEU and TER, score translations largely based on n-gram overlap, the simple loss of a pronoun has little impact. In a prior paper, we demonstrated that this tendency leads to translations with a lower density of referential cohesion, and an according reduced comprehensibility – an effect which is amplified in literary texts and other non-newswire domains (Voigt and Jurafsky, 2012).

Therefore, in this work we aim to propose a novel technique based on pivoting that will allow for the recovery of dropped arguments, hopefully in the end improving MT quality and comprehensibility.

2. Linguistic Analysis

Before engaging in a study of this sort, it is prudent to develop an understanding of the current consensus in linguistics as to the status of null arguments in Chinese. In reviewing the literature, I found that the structural analysis of these phenomena is very complex, and there is not broad agreement among linguists as to their structural role. For the most part, linguists propose that null subjects are straightforward subject-position pro-drop (a.k.a. *pro* or “little pro”), analogous to the subject drop we find in languages such as Italian and Spanish (Huang 1989); however, such an analysis is complicated by the fact that such a proposed *pro* in Chinese lacks an overt structural representation of pronoun-verb agreement, a feature postulated to be obligatory for the type of pro-drop seen in subject drop languages.

Null objects, on the other hand, are more complicated, and have been interpreted variously as instances of pro-drop, A-bar-bound variables, VP ellipsis, and NP ellipsis (Huang 1991, Otani and Whitman 1991). The most contemporary, and to my mind most convincing, accounts propose all phrasal ellipsis in Chinese to be instances of a “true empty category” licensed by subcategorization features, with a case-fulfillment restriction for nominal phrasal ellipsis as in object drop (Aoun and Li 2008).

The complications involved with the analysis of object drop in Chinese motivate our first experiment, that is, they motivate an explicit handling of subject *pro* but ignoring instances of object argument drop for the time being.

3. Experiment 1: *pro*-identifying Classifier

Previous computational linguistics work in this area has focused on identifying and resolving instances of subject-drop *pro* automatically by means of syntactic features extracted from small amounts of training data (Zhao and Ng 2007, Kong and Zhou 2010). Given the complexity of the phenomena involved with object drop, and given that subject drop is a far more prevalent phenomenon, we put object drop aside and began a first experiment similarly attempting to identify subject *pro* in the style of Zhao and Ng (2007). Though Kong and Zhou also engaged in similar work, they did not respond to requests for access to the dataset they used, and so we would have no basis for meaningful comparison with their work. Furthermore, it is unclear from their work that they were, in fact, trying to find instances of *pro*, but instead they appear to also be

looking for non-expressed features such as NP-traces and variable binding that will have no linguistic expression and thus that would be no use to an MT system.

We hoped that building such a classifier, if successful, would allow us to improve translation quality by either: a) retraining an MT system with an inserted novel token such as '*pro*' in the training data that would be aligned with English-language material during alignment and phrase extraction and then, at decoding time preprocess the source again with inserted '*pro*'s and allow the system to choose an appropriate translation; b) inserting the most common pronoun, such as *ta* the male/neuter pronoun, in zero locations and doing translation; or c) later building a second classifier to classify *pro* locations with their necessary antecedent.

Thus to begin we made a re-implementation of Zhao and Ng's zero anaphora identification classifier, which is trained and tested on trees in the Chinese Treebank. In their work they generate zero pronoun (ZP) candidates by simply identifying all gaps between words to the left of some VP. We used Stanford CoreNLP's linear classifier with conjugate gradient ascent, and implemented features analogous to theirs. These can be found in detail in their paper, but they include features such as: Is this the first ZP gap in the sentence? Is this gap adjacent to a comma? What is the lowest shared node in the parse tree between the words to the left and right of this gap? What is the parse structure of each of these nodes? In my code these features are labeled with comments analogous to how they are described in the Zhao and Ng paper, and should thus be easily identifiable.

4. Experiment 1: Results

Our initial experiments proved successful relative to Zhao and Ng. We contacted them and they were kind enough to provide us with access to their data, which allowed for direct and meaningful comparison. This success is perhaps largely attributable to the use of an improved classifier more appropriate to the task as compared with the Weka decision tree classifier (Daniel Cer, personal communication).

Zhao and Ng were using CTB 3.0, and the Chinese Treebank is now at version 7.0, so we also included tests where we trained on the entirety of the newswire section of CTB 7.0. We found that training on the entirety of the Chinese Treebank, which now includes weblog and other cross-domain data, had an adverse impact on performance, an issue that will become important later in this paper.

Zhao and Ng were training on only 155 documents and testing on 50 documents: to establish a new, modern baseline with the full Treebank, we used the standard split suggested by the Treebank authors to train and test our classifier. Furthermore, in anticipation of the necessity of greater robustness for MT tasks, we took the raw sentences from the Chinese Treebank and automatically parsed them with the Berkeley Parser, and got results for our system trained on those parsed sentences. All the aforementioned results are listed in the table below:

Table 1. Experiment 1 Classifier Performance at Identifying *pro*

Training Data	*Test Data*	*Precision*	*Recall*	*F1*
Zhao and Ng (2007) Best Results		59.8	44.3	50.9
Zhao and Ng Train	Zhao and Ng Test	62.8	56.3	59.4
Treebank Newswire	Zhao and Ng Test	66.2	56.3	60.9
Treebank Gold Training Split	Treebank Gold Dev Split	69.4	58.5	63.5
Treebank Berkeley-Parsed Training Split	Treebank Gold Dev Split	56.5	60.1	58.2

5. Cool-to-Hot Pivot Proposal

In developing the classifier described above, we noticed several points. First of all, the classifier is very fragile to data in new domains. Second, the performance is not sufficient to necessarily improve MT performance – if zeroes from a system with an F1 in the ~60 range are inserted, its unclear if it won't even hurt the quality of the output. Finally, prior work has focused on high-quality gold-parsed Treebank data, and MT systems in general must be far more robust to new inputs, and do not generally have the luxury of operating on such clean data.

Therefore, in this work we propose a novel technique inspired by prior work in the MT community using “pivoting” to improve machine translation systems between languages that do not have large bilingual corpora, as a means of offering functionally “infinite” training data for this task. In prior “pivoting” work, researchers leverage an intermediary language for which large bitexts exist with the other two languages to increase overall MT performance (Wu and Wang 2007). Here, we propose a two-language “pivot”-inspired approach, in which we extract additional information from MT bitexts that

is not captured by traditional phrase-based MT systems.

First, remember the discussion from the introduction regarding a proposed “hot-cool” parameter. “Hot” languages more explicitly express their anaphoric elements, while “cool” languages utilize ellipsis, argument drop, and other forms of gapped material that obligate greater inferrational work on the part of the hearer. We begin from the intuition that in the large corpora of dual-language translated bitexts used to train machine translation systems, when the source language is "cool" and the target language is "hot,” translators have necessarily done massive amounts of pragmatic disambiguation of null arguments and other "discourse-oriented" features.

We then aim to leverage this intuition to extract this disambiguated material as training data for our classifier to increase its accuracy and robustness. We propose the following procedure for identifying dropped subject pronouns:

- Given a large translated bitext, perform high-quality word alignment.
- Identify unaligned subject pronouns in the “hot” language.
- Observe the words aligned to their left and right, and identify the span between the words in the “cool” language to which they align.
- If this span has a length of 0, propose it as a heuristically identified zero (HIZ) and thus a positive training data example.

A visual representation of the algorithm can be seen in the following example:

Figure 1. Visual representation of the heuristic zero identification algorithm. 'he' is the unaligned pronoun, lines between words represent alignments, and the dotted blue line represents the insertion proposal.

We tested this procedure on various standard MT datasets for Chinese-to-English translation (the Stanford MT research group's internal “grammatical subset,” splits from the BOLT project bitexts, and so on), and found that in general 15% of sentences in the bitexts contained subject-position HIZs, and for 25% of those the source-side span was of length 0. This may seem at first to be quite limiting, considering that this means only 4% of sentences may be proposed as containing positive training-example HIZs. However, note that in a 10-million sentence bitext of the type commonly seen in MT system training, this results in a total of 400,000 positive training examples. Compare this with

Zhao and Ng, who in their work trained their classifier on only 665 positive training examples.

6. Experiment 2: Parallel Treebank *pro*/HIZ Comparison

To test if heuristically identified zeroes as generated in the algorithm described above indeed match up with what linguists identify in formal datasets as instances of subject-drop *pro*, we performed an experiment testing their degree of overlap with *pro* in the Chinese Parallel Treebank, a subset of documents from the Chinese Treebank with corresponding English translations.

We identified and inserted potential zeroes into this dataset using our cool-to-hot pivot technique, and found these zeroes to have a precision of only 35% in terms of their overlap with *pro*. This was at first a surprising finding, but looking manually at the data revealed certain patterns. First, zeroes we inserted were often either adjacent to Treebank *pro* or, perhaps more crucially, were not labeled as zeroes at all in the Treebank but still appear that they would be useful for MT applications. Observe the following example:

Gold: 丁豪 在 儿童 福利院 读完 小学，
Hao-ding in child welfare-institutiton read-finished elementary-school
随后 进入 附近 乡里 一 所 学校 上 初中。
then enter area township one M school go-up middle-school
Ref: Hao Ding finished primary school at the children's welfare institution, then he entered a school in a nearby township to go to middle school.
Test: 丁豪 在 儿童 福利院 读完 小学 ， 随后 *zero* 进入 附近 乡里 一 所 学校 上 初中 。

In this example we note that our heuristic algorithm proposes a zero between “then” and “enter” in the source Chinese, precisely the location where the English translator inserts a “he” to offer a more cohesive and fluent translation in the English.

This experiment with the Parallel Treebank strongly suggested that identifying “true” linguistic zeroes, defined as *pro* annotated by expert linguists in the development of the Chinese Treebank, was in fact the wrong task, and would not provide appropriate training data for a *pro*-identifying classifier. To confirm this suspicion, I implemented an ad-hoc capability in the classifier described above to read berkeley-parsed bitext trees as training data, and anecdotal tests with 400,000 parsed trees found abysmal performance as expected, with an F1 score hovering around 10% when testing on the Treebank dev set. It seems clear that this result is derived from a combination of the mismatch between

our heuristically identified zeroes and “true” linguistic zeroes with an aggravated domain-specificity problem: the MT-derived training data is inherently much more noisy and broad in terms of domain.

7. Refocusing the Task: Translator-like Insertion with Cool-to-Hot Pivoting

The combination of our results from our first two experiments made clear that, in fact, if we want to work towards an applied goal of improving MT performance, finding subject-position *pro* is the wrong task. Not only do linguistically identified *pro*s not match up with zeroes identified by our proposed cool-to-hot heuristic insertion algorithm, but it seems to be the case that the zeroes that our algorithm does find are perhaps more appropriate for the MT task than “true” linguistic *pro* in Chinese.

Therefore, we propose a new task: build a classifier that identifies locations in a given Chinese source text where a translator would likely choose to insert pragmatic materials in a translation that are not overtly present in the source. By pragmatic materials, we mean to capture the intuition from the linguistic analysis in section 2 that many “cool” or “discourse-oriented” linguistic features can be described as phrasal ellipsis subcategorized by some material in the text.

One unfortunate consequence of this shift in task is that it reflects a far more utilitarian and less principled approach. While *pro* is a relatively theoretically well-defined phenomenon, 'places where a translator inserts something' has a far shakier theoretical grounding. However, we find some principled support for this idea in earlier empirical studies that demonstrate the effectiveness of, essentially, “English-izing” Chinese for the sake of improved MT (Wang, Collins and Koehn 2007; Chang et al. 2009), and it is this concept that our new approach captures.

All in all, this is an exciting turn for this research: it implies that not only may we be able to improve MT performance with anaphoric elements such as subject and object drop, but we may also be able to improve performance with VP ellipsis, conditional phrases, pleonastic 'it' in English, and other discourse features Chinese does not overtly express. Furthermore, formulating the task in this manner allows for the possibility that our approach will prove widely applicable to many language pairs, not just Chinese-English, where the source language is somewhat more “cool” than the target.

8. Data Annotation

However, one disadvantage of this reformulation is that given we are now dealing with a completely new task, there are no available datasets on which to test the performance of any system we might build. Therefore in the current stage of the project, I am in the process of manually annotating several standard machine translation datasets to act as appropriate test sets for what will become our new classifier. Currently we are only annotating these datasets with pronominal zeroes, specifically 我 (*wo,* 'I/me')，我们 (*women*, 'we/us')，他 (*ta*, 'he/him')，她 (*ta*, 'she/her')，他们 (*tamen*, 'they/them'), although as mentioned beforehand later passes through the data may involve annotation of additional features relevant to the task.

In order to establish a principled methodology for this annotation, we employ the following guidelines:

- Only annotate sentences for which more pronouns occur in some English reference translation than in the original Chinese source
- In such sentences, where a translator has inserted a pronoun, annotate that pronoun in the most “English” position, that is, if word-by-word glosses of the Chinese text were provided to an English speaker, where they would find its insertion most natural
- Annotate with a #Z tag so that zeroes can later be manipulated, replaced, and identified as distinct from other possible discourse-annotated information

9. Experiment 3: Google Translate Oracle

Thus far, we have used some of this annotated data to perform proof-of-concept oracle experiments with Google translate. These basic experiments hope to confirm our intuition that in fact this style of insertion is the appropriate course of action. They also represent the most basic possible formulation of our planned eventual classifier: one that annotates *only* the source-side text at decoding time, without retraining the system or otherwise interacting with its internals in any way. In a real experiment down the road we would expect increased performance from these deeper adjustments.

Table 2. Experiment 3 Oracle BLEU scores using Google Translate. 'Original' refers to Google-translated versions of the unannotated source, and 'Fixed' refers to Google-translated versions of the source with gold pronoun insertion annotations.

Dataset	*Original BLEU*	*Fixed BLEU*
Chinese Parallel Treebank, 1000 sentences (38 annotations)	18.639	19.172
GALE DEV12 weblog dev, 300 sentences (91 annotations)	15.851	16.387

For both datasets, we find a gain of approximately half of a BLEU point. This is not a stellar result, especially considering that it is based upon gold annotations, a level of accuracy we cannot hope to achieve with a classifier. However, keep in mind that these results are on a non-retrained system, using an n-gram based automatic metric that has no conception of argument structure. Looking by hand at some of the results, it seems to be the case that in some instances correctly inserted pronouns coerce the system into providing better contextual translations overall, as in the following example from the Treebank:

Original: 看来 ， 还 要 多 在 室外 训练 。 ”
Fixed: 看来 ， 我们 还 要 多 在 室外 训练 。 ”
Ref: Looks like we still need to train more outdoors."
Google Original: It seems also to outdoor training. '
Google Fixed: It seems that we need more outdoor training. '

Notice also that there were a higher proportion of sentences requiring annotations in the weblog data (30.3%) than in the Parallel Treebank newswire data (3.8%), suggesting that the effects we are observing are highly domain specific and likely much more visible in more colloquial non-newswire text.

One last important piece of data from this oracle experiment regards the distribution of zeroes. In the DEV12 weblog data, we found the following distribution:

wo	21
women	34
male *ta*	1
female *ta*	9
tamen	26

This is striking. Whereas we had previously expected zeroes to largely take the form of male *ta* ('he/him'), in the weblog data they are widely distributed. In annotating

the data I had noticed that the distribution seemed to be the result of the domain. For example, *wo* ('I/me') was prevalent in forum and blog posts, *women* ('we/us') was prevalent in discussing large-scale ideas for what 'We the Chinese people' ought to do politically, and *tamen* ('they/them') was used in discussing the actions of political actors such as nations. This result suggests that though we cannot ignore resolution after zeroes are identified, their contexts may be very specific and thus amenable to high-accuracy resolution with a classifier.

10. Conclusions

While this research has not yet been able to produce a full-fledged zero anaphora detection and insertion system that can improve machine translation quality, I feel strongly that the work presented here offers promising inroads to solving this difficult problem. Moving forward, we hope to use our heuristic identification algorithm to produce a high-performing classifier that can be demonstrated to improve Chinese-English MT performance, both by inserting zero markers or their referents in training data and by discovering them explicitly at decoding time.

Though oracle experiments as described above show modest BLEU gains, it is very possible that demonstrating the effectiveness of such an approach will require the use of a more complex automatic evaluation metric. Lo and Wu (2011) have presented a promising metric entitled 'MEANT' which uses semantic role fillers to evaluate argument structure, and we are currently in the process of trying to obtain a research version of it for this purpose, since presumably our approach would do demonstrably better measured on such a metric. We hope this work will provide impetus for further research on handling pragmatic and discourse-level considerations in machine translation.

References

Joseph Aoun and Audrey Li. 2008. Ellipsis and missing objects. pp. 251-274. MIT Press, Cambridge: Foundational Issues in Linguistic Theory/MIT Press.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. Discriminative Reordering with Chinese Grammatical Relations Features. In Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation.

C.-T. James Huang. 1987. Remarks on Empty Categories in Chinese. Linguistic Inquiry 18:321-337.

C.-T. James Huang. 1991. Remarks on the status of the null object. In Principles and parameters in comparative grammar, ed. Robert Freidin, 56-76. Cambridge, MA: MIT Press.

Fang Kong and Guodong Zhou, 2010. A Tree Kernel-based Unified Framework for Chinese Zero Anaphora Resolution. In EMNLP.

Chi-kiu Lo and Dekai Wu. 2011. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In ACL.

K. Otani and J. Whitman. 1991. V-raising and VP ellipsis. Linguistic Inquiry 22: 345-358.

Rob Voigt and Dan Jurafsky. 2012. Towards a Literary Machine Translation: The Role of Referential Cohesion. NAACL Workshop on Computational Linguistics for Literature.

Chao Wang, Michael Collins and Philipp Koehn. 2007. Chinese Syntactic Reordering for Statistical Machine Translation. EMNLP.

Hua Wu and Haifeng Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. Machine Translation, 21(3): 165-181.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach. In Proceedings of EMNLP CoNLL Joint Conference.