CS 224N

Using Named Entity Recognition to improve Machine Translation

Neeraj Agrawal

Ankush Singla

Abstract

Named Entities (NEs) are a very important part of a sentence and its important for a Machine Translation (MT) system to get them right. Mistranslating or dropping NEs may not change a sentence's BLEU score by much, but it can hurt sentence's human readability considerably. Current MT systems consider NE's as normal words and subsequently drop or mistranslate them during the translation. In our paper, we present three different approaches to help these systems treat NE's differently and help improve the human readability overall.

Introduction

Our focus is on Chinese to English translation. We start with the current Stanford Machine Translation system. Treating the Stanford system as a reference standard, we experiment with using Named Entity Recognition (NER) for improving its performance. We experimented with three different ideas of using NER data: adding named entity data to help the aligner, modifying the language model, and giving higher weights to translations with same number of Named Entities (NE) in source and target. We discuss each of these ideas in detail, algorithms used and the motivation for choosing them. We also provide an in-depth analysis of the performance, highlighting the improvements as well as point out the cases where performance decreased.

Previous Work

Most of the published work that uses NER for MT has been directed towards learning to transliterate NEs. The work done by Ulf Hermjakob and Kevin Knight et. al. [1] for Arabic-English translation shows that improvement in MT can be achieved by transliterating NEs in source data (instead of trying to translate them). It builds on the hypothesis that MT system drops or mistranslates NEs when they do not occur in the training data.

Another paper from Bogdan Babych et. al. [2] uses a simpler approach for European languages. It creates a "Do Not Translate" list based on NER and simply copies any word in this list from source to target without any translation or transliteration. Their analysis shows that 'well-formedness' of sentences improves by this simple approach.

Performance of Current MT systems on NER

Current MT systems tend to drop or mistranslate NEs in a sentence, which doesn't hurt the BLEU [6] score much but it does hurt a human's readability. We found many examples of this in the output from current Stanford MT system as well:

Chinese: 27日 中午 , 他们 已 被 安全 转移 到 普吉岛 。
English: 27 noon , they have been shifted to safe places to 3pm .

Here Chinese phrase '普吉岛' means 'Phuket Island' but MT system dropped it from the translation. The available reference translations also show that we should have 'Phuket Island'. Similar observation was made in a number of other sentence translations.

Chinese: 据 泰米尔纳德邦 首府 钦奈 的 渔业界 权威 人士 介绍 , 这 . . . .
English: "the authority of the capital , according to Nadu fishery source . . . .

In this sentence, system dropped 'Tamil' from the translation of '泰米尔纳德邦 ' which means 'Tamil Nadu'.

Experiment 1: Adding NER data to help aligner

One way to help the aligner is to append the list of NEs to the bi-text training data. This will increase the probability of matching these NEs in source language to their counterparts in target language. It will also increase the vocabulary of the target language (these NEs may not have been present earlier in the training sentences).

For this purpose, we used the named entity data provided by Linguistic Data Consortium [7]. It contains a list of Chinese-English bi-directional NEs compiled from Xinhua News Agency newswire texts. We preprocessed this data to match the original bi-text training data. Pre-processing included changing encoding of Chinese text from GB to UTF8 and adding spaces before and after special symbols (+ , - , / , ; , . , \ ).

The following BLEU [6] scores were obtained:

	Dev Set	Test Set
Base Case	31.968	31.349
With NER Data	31.916	30.268

Analysis

1. BLEU score for development set doesn't change much. However, BLEU score for Test set decreases considerably (approx. 0.9)

2. Analysis of individual sentences reveals that this approach does improve performance for NERs that were present in LDC data. Consider the example:

   Chinese: 27日 中午 , 他们 已 被 安全 转移 到 普吉岛 。
   Translation: at noon on the 27th , and they have been shifted to safe places to phuket .

   Chinese: 据 泰米尔纳德邦 首府 钦奈 的 渔业界 权威 人士 介绍 , 这 场 突如其来 的 . . . . .
   Translation: "according to an authoritative source , acting for the state of tamil nadu capital . . .

   Here we can see that whereas base case had dropped 'Phuket' from the sentence, adding additional data helped it to align the Chinese word 普吉岛 to Phuket. Similarly for second sentence it preserved 'tamil nadu' whereas base case had dropped 'tamil'.

3. However, there were also examples where translating NEs reduced BLEU score. For example:

   Chinese: 法新社 伦敦 六日 电
   Base Case: hong kong presse , 6 (xinhua)
   NER Translation: agence france presse london 6 (xinhua)
   Ref 0: afp on the 6th , london
   Ref 1: afp , london , 6th .

   Here we can see that 法新社 was originally translated to 'hong kong presse', whereas after adding additional data its correctly translated to 'agence france presse'. Its also noteworthy that while this is correct, it doesn't help in BLEU score as all reference translations had the corresponding translation as acronym "afp".

Improvement based on analysis

Analysis of the NEs list provided by LDC showed that although most of the foreign words are English or can appear in English texts, there are also many non-English words. This was also observed in the example above where 'afp' was written as 'agence france presse'. We therefore made an effort to refine the original NEs list. We used the following filtering:

- For lists containing names of industry / organization / press only those NEs were kept that belonged to UK, USA, china, UN.
- For lists containing Chinese proper names, no filtering was done.
- List containing place names was discarded as it had a large number of non-English words.

The following BLEU scores were obtained:

	Dev Set	Test Set
Base Case	31.968	31.349
With NER Data	32.276	30.427

Analysis

1. An improvement of approx. 0.3 was observed in the development set BLEU score over base case. The BLEU score for test data improved over using complete list of NERs. However, it was still lower than base case.

2. Translations contained less of non-English words. Consider the same example discussed earlier:

   Base case : hong kong presse , 6 (xinhua)
   All NER data: agence france presse london 6 (xinhua)
   Filtered NER data : afp london , 6 (xinhua)

   Here, filtered NER gave the translation 'afp' as in reference sentences, hence had higher BLEU score.

3. The system failed for sentences containing NERs that were not present in the training NER list:

   Base case: bush has also admitted that the training results of iraqi security forces . . . .
   Filtered NER: interim national security Seventy - buhi also admitted that the . . . .

4. Performance over sentences with no NER seems to have decreased in general:

   Base case: i 'm still very enthusiastic about this sport . "
   Filtered NER: i still very enthusiastic about this movement . "

   Base case: earlier the state council released the contents of the interview .
   Filtered NER: earlier in the day with the contents of the interview released by the state council .

   Here, the reference sentences contain no NER. It seems that the base case gave better 'formed' sentences than the one with additional NER data. Observing the weights for language model (LM) in each case revealed that weight for LM is higher in base case (0.053) than in the latter (0.04), which explains the better sentence formation in base case in the absence of NEs.

5. Another interesting observation is that the average translated sentence length is considerably smaller in base case than in the case with added NER data.

   Base case avg. sentence length = 33.8
   NER data avg. sentence length = 34.8

   This also supports the observation that base case drops NEs during translation.

Experiment 2: Class Based Language Model

In the current language model, we can get very low probability score for a sentence with NEs just because we had never seen those NEs in our training data. For example, if we have seen 'David' appearing a lot of time in our training data and we haven't seen 'Ankit' a lot, then the following two sentences will have different LM scores:

"David is going for a walk"
"Ankit is going for a walk"

Here the first sentence will get higher score in comparison to first. This can reduce performance since the test set usually contains a lot of previously unseen NEs. Our idea here is to replace all the person names by a token 'PERSON' and all the organization names by 'ORGANIZATION'. This will make sure that LM model will not reduce score for a sentence just because it has a name that was never seen in training data.

Step by step Implementation:

1. First of all, we run Stanford NER system [3] on Language model training data and replace all person names with 'PERSON' and organization names with 'ORGANIZATION'.

2. After training LM, with this data we need to make sure that we do the same processing for every sentence before finding its LM score. However, running an NER on each hypothesis has the following issues:

   - Running NER on each hypothesis will be really expensive.
   - We require a complete sentence to run NER, (as some NER features depend on next words) which may not be available during decoding.

3. To work around these issues, we changed the way words are stored in phrase table itself. We ran NER on the English sentences of the bi-text training data given to translation model and changed all names to include class label information
   e.g. 'David' to 'David/PERSON' .

4. This will give us phrase table with words like 'David/PERSON' and 'Stanford/ORGANIZATION' and we can simply use the second part of the phrase to find LM score. So before finding LM score we replaced 'David/PERSON' to 'PERSON'.

5. Finally we had to replace David/PERSON with David before obtaining BLEU scores during MERT [4] (Minimum Error Training) as our references are not NER tagged.

The following BLEU scores were obtained:

	Dev Set	Test Set
Base Case	31.968	31.349
With class based LM	26.935	25.532

Analysis

1. The BLEU score for both development and test set decreased in comparison to base case.

2. The model performed better for certain sentences containing a number of PERSON / ORGANIZATION tokens. Consider the example below:

   Base case: but the ball still insisted that the . . .
   Reference: however , powell still firmly holds . . .
   Class based: however , powell still insisted that the . . .

   Here, our model correctly translated the sentence to contain the phrase 'powell' , whereas the base case translated the corresponding Chinese phrase into 'ball'. The word 'powell' appears very rarely in LM training data and hence is mistranslated by base case model.

3. The model added a number of person / organizations even when not present in the original Chinese sentence. Consider the examples below:

   Base case: at that time , and with the support of about 30 people , shook hands , that is , vomiting , chairman of the phenomenon .

   Reference: " the president shook hands with around 30 supporters before leaving to vomit .

Class based: At that time , about 30 people support President Jiang and Deputy President Jiang shook hands with , that is , vomiting .

Here, the current model added word 'Jiang' twice (tagging it as a 'PERSON') whereas it never appeared in the reference translations. This is happening because our LM had a very large counts for n-grams containing 'person' or 'organization' tokens. Another example for the same is:

Base case: johnson , however , said that due to . . .
Reference: however , johnson says that . . .
Class based: However , the Johnson & Johnson , indicated that

Here, the reference translation contains person 'johnson'. Our model, however, prefers the translation 'johnson & johnson'. Analysis revealed that it tagged each of the 3 words as ORGANIZATION and hence the class based LM model gave it a relatively higher score as discussed above.

4. Overall, the performance decreased for a large number of sentences. We suspect that this happens because using class based LM model is too severe an approach: The tags 'PERSON' and 'ORGANIZATION' have been given a much higher count than most other word in the LM. We therefore expect a linear combination of class based and normal LM to perform better. While the class based could help in obtaining a non-zero score for names that do not occur in training data, the normal LM would ensure that names aren't given exceptionally high score either.

Experiment 3: Adding Extra Feature

In a Chinese – English translation, we will expect a correct translation to have equal number of NEs in both source and target. However, there is no feature in the current Stanford MT system that considers NEs differently than any other words. We therefore introduce a new feature that favors phrase translations with equal number of NEs.

Implementation

Our overall goal is to run NER on the source sentence and on all target sentence hypotheses during decoding. We could then increase the weights for hypothesis with equal number of NEs. However, there were two main issues with this approach:

a. We do not have a good Chinese NER,
b. Running NER on each hypothesis would be very expensive.

In the absence of a good NER system, we decided to use counts of proper nouns instead of NEs. We used the available Stanford POS tagger [5] to get the proper noun counts. The argument that the source and target language should have equal number of NE's holds true for proper nouns as well. This helped us solve the second issue as well, as running POS tagger is not as expensive as NER.

Feature Value = exp (abs (Difference in number of proper nouns in source and target))

The following BLEU scores were obtained:

	Dev Set	Test Set
Base Case	31.968	31.349
With extra feature	32.155	30.914

Analysis

1. An improvement of approx. 0.2 was observed in the development set BLEU score over base case. However, the BLEU score of test set decreased by 0.4

2. The additional feature is given a negative weight by MERT (-0.03), which confirms our hypothesis that it prefers sentences with lower difference in number of proper nouns across source and translation.

3. Phrases with same number of proper nouns as in source are preferred. Consider the example:

   Chinese 巴林 公主 下 嫁美 大 兵惊世 婚姻 五 年 宣告 破裂
   Base case: bahrain declared in the next five years , the princess big marriage break-up
   Extra feature: princess of bahrain next big marriage broke down in five

   Here, the Chinese phrase '巴林 公主' means 'Bahrain Princess'. This is correctly translated into 'princess of bahrain' by current model with extra feature. However, in the base case this was split into separate phrases. Our model gives higher score to the former phrase as it had equal number of proper nouns as corresponding Chinese phrase. The same trend was also observed in a number of other sentences.

4. A couple of other interesting examples are:

   Chinese: . . . 中国 总理 . . .
   Base case: chinese prime minister
   Extra phrase: chinese premier
   Reference : chinese premier

   Base case phrase: johnson and johnson said : " she has been to the end . "
   Extra phrase: johnson said : " she has been to the end . "
   Reference: " she 's gone off the deep end , " johnson says .

   In the first example, the Chinese phrase has two proper nouns. Hence, our model with extra feature prefers the phrase 'chinese premier' which has 2 proper nouns over 'chinese prime minister' which has 3 proper nouns. The base case model however preferred the phrase 'chinese prime minister' giving a different meaning to the sentence.

Likewise, in the second example, 'johnson and johnson' is preferred by the base case model. Our extra feature ensures that 'johnson' phrase is selected thereby leading to a better translation.

5. The weight given to LM by MERT decreased from 0.053 to 0.039. This distorted the formation of some sentences. Consider the examples below:

   Base case: nas said that pregnant women should still plan taking folic acid supplements .
   Extra feature: nas said that the plan of the pregnant women should still taking folic acid supplements .

   Here, we expect the trigram 'should still taking' to have a low LM score. However, other features may have dominated in our model as LM was given a relatively low weight. The base case model chose 'still plan taking' as this trigram is more likely in training data and thus has a higher LM score.

Conclusion

In conclusion, it can be said that using Named Entities does help in providing better Machine Translations. Although, we could not improve the BLEU score over current Stanford MT system, a number of translated sentences did show improvement with the use of these techniques. There was a considerable reduction in dropping and mistranslation of NEs. This helped enhancing human readability as well. Analysis of our models also revealed a number of insights and scopes for further improvement.

Future Work

1. The NE list obtained from LDC is not well processed. Words are not always capitalized and it contains certain non-English words. We also saw that our results improved after doing some processing on the data. Further processing / filtering might help us improve the results even more.

2. We used proper nouns in the third approach (of adding extra feature) in the absence of a Chinese NER. It would be interesting to see the performance with NER counts instead of POS.

3. We might have been too aggressive in our approach in using a class based language model. An alternative to this could be to use a linear combination of normal LM and class based LM.

4. We would also like to try different combinations of these three approaches. We couldn't do it here because of time constraints.

Acknowledgements

We would like to extend our special thanks to Professor Christopher Manning, Daniel Cer and Michel Galley for their valuable suggestions and feedback.

References

1. U. Hermjakob, K. Knight, and H. Daume III, Name Translation in Statistical Machine Translation: Learning When to Transliterate, Proc. ACL, 2008.

2. Babych, Bogdan; A. Hartley. Improving machine translation quality with automatic named entity recognition. 2003. In Proceedings of EAMT/EACL 2003 Workshop on MT and other language technology tools.1–8. Budapest, Hungary.

3. Jenny Rose Finkel, Alex Kleeman and Christopher D. Manning. 2008. Efficient, Feature-based, Conditional Random Field Parsing. Proceedings of ACL/HLT-2008, pp. 959-967.

4. Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In ACL 2003: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics.

5. Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

6. Kishore A. Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY, September.

7. Shudong Huang 2005 Chinese <-> English Name Entity Lists v 1.0 Linguistic Data Consortium, Philadelphia