

Parsing Statistical Machine Translation Output

Simon Carter and Christof Monz

University of Amsterdam
Science Park 107, 1098 XG Amsterdam, The Netherlands
{s.c.carter, c.monz}@uva.nl

Abstract

Despite increasing research into the use of syntax during statistical machine translation, the incorporation of syntax into language models has seen limited success. We present a study of the discriminative abilities of generative syntax-based language models, over and above standard n-gram models, with a focus on potential applications for Statistical Machine Translation (SMT). We show that in fact parsers are better able to discriminate between good and bad English, and that parsers, as well as n-gram language models, assign higher average log probabilities to references in comparison to SMT output.

1. Introduction

The fundamental equation of machine translation formulates Statistical Machine Translation within the noisy channel framework:

$$p(e|f) = \operatorname{argmax}_e p(f|e)p(e) \quad (1)$$

The translation model is given by $p(f|e)$, and $p(e)$ represents a language model. The most commonly used language models (LM) are word n-gram models, which base the probability of a word following a prefix string on a limited context of the previous $n - 1$ words. Given the nature of n-gram models, there is a strong motivation for the applicability of syntactic LMs that can take into account long-range dependencies that are difficult to model by the limited context of n-gram models.

Yet the standard n-gram LMs have seen limited progress in their development from the n-gram based models first used in speech. Attempts to integrate fuller syntactic information into Language Models used by Statistical Machine Translation systems has met with mixed success.

Och et al. (2004) highlight the problems that syntactically informed LMs can have; they report the average log-probability assigned to SMT output by a parser to be higher than that assigned by the same parser to the reference sentences. Surprisingly, the parser assigned a higher probability to ungrammatical SMT output than supposedly good, human written sentences.

One potential explanation for the limited success of parsers as language models is the function for which they were originally designed. Parser accuracy is a measure of how much a test parse matches a given gold standard. However it is not clear how parsers react with sentences where assumptions about its grammaticality and fluency do not hold. A major motivation for our work is thus to see to how parsers operate over examples of ungrammatical natural language.

To this end, we design an experiment to compare different parsers on their ability to discriminate between examples of well-formed and ill-formed English, trained on various data sets.

The remainder of this paper is organized as follows: In Section 2 we discuss related approaches which motivate our own research. In Section 3 we introduce the parsing models used in our experiments, including our own model; a Lexicalized Expectation model. In Section 4 we present out results, discuss them in Section 5 and finally draw conclusions in Section 6.

2. Related Work

Charniak et al. (2003) use a head-driven language model coupled with a syntax based translation model. They also show promising results, though there is a trade-off between grammaticality and semantic correctness. Och et al. (2004) unsuccessfully try a number of different approaches to n-best re-ranking, including numerous different syntactic approaches.

More recently, the approach of Post and Gildea (2008) decouples the isomorphic relationship existing between syntactically motivated translation rules and target side grammars, however they are unable to outperform the baseline.

3. Syntactic Parsing

Probabilistic parsers assign a syntactic structure to a given sentence, with an associated probability. A probabilistic context free grammar (PCFG) G is defined by five parameters (N, Σ, P, S_t, D) , where N is a set of non-terminals representing syntactic constituents, Σ the set of words, P a set of productions, S_t a designated start, or root symbol, and D a function for assigning probabilities for each rule in P .

A rule is thus defined to be in the form $\Lambda \rightarrow \alpha p$, where Λ is a non-terminal, and can be said to expand into α , which is the set of symbols from the set $(\Sigma \cup N)$. Λ will be referred to as the *LHS* and α the *RHS*. The probability associated with that rule is represented by p . The probability of a parse tree is simply the product of all rules used in the parse.

3.1. Earley Parsing

In exploring the use of parsers as language models for stack-based decoding, we developed a parser according to

| Model | All Sentences | | | | Length ≤ 40 | | | |
|------------|---------------|---------|--------|------------------|------------------|--------|---------|------------------|
| | LR | LP | F1 | Tagging Accuracy | LR | LP | F1 | Tagging Accuracy |
| PCFG | 64.53 % | 64.29 % | 64.41% | 93.30% | 65.76 % | 65.70% | 65.73 % | 93.29 % |
| Collins M2 | 86.07% | 87.82% | 86.95% | 94.44% | 86.56% | 88.44% | 87.5% | 94.28% |
| LE | 55.66% | 65.59 % | 60.22% | 40.38% | 56.73 % | 67.25% | 61.54 % | 40.08% |

Table 1: Labeled Precision and Recall scores for the different parsers, trained on sections 2–21 and tested on section 23 of the Wall Street Journal Penn Tree Bank

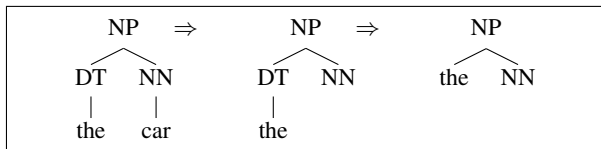


Figure 1: An example showing the extraction and transformation of rules into Greibach Normal Form.

the Probabilistic Early framework, as defined by Stolcke (1995). We chose the Earley framework over the more commonly used CYK approach because it allows one to easily compute probabilities over prefix substrings, generating parse trees in a left to right manner.

This matches the way most SMT systems generate hypothesis, allowing us to take into account the syntactic probabilities associated for potential translations alongside the other models in the log linear translation framework. There is also the opportunity to streamline the parsing workload by sharing charts across those hypothesis which share a parent state, continuing to parse only the newest part of the now extended hypothesis.

3.2. Basic PCFG

We include a basic PCFG parser in our experiments for comparison purposes. We take a ‘basic’ PCFG to be one in which the productions in our grammar are either non lexicalized, where a non terminal (NT) token expands into a set of NTs, or a NT expands into a single word.

3.3. Collins Model 2

We used Dan Bikel’s implementation of Collins Model 2 (CM2) (Bikel, 2002) as a state-of-the-art parser to be used as our baseline. Collins Model 2 assigns four different probability distributions for assigning probabilities to (i) head labels, (ii) sibling labels and head tags, (iii) sibling head words, and (iv) subcategorization frames. We chose Model 2 above Model 1 because of the higher reported Labeled Recall and Precision scores (Collins, 1997).

3.4. Lexicalized Expectation Model

A problem encountered in (Post and Gildea, 2008) is the production of good syntactic structures over disfluent sentences. To help overcome this limitation, we constructed a grammar in the Greibach Normal Form. Rules in Greibach Normal Form are either in the form $A \rightarrow a\lambda$, where A is a NT, a is a word, and λ is a set of NTs, which can be empty. Thus all rules have a lexical item, coupled with a specific set of NT expectations. Due to the explicit lexicalization within the grammar, we hope to construct syntactically correct parses over good sentences, while at

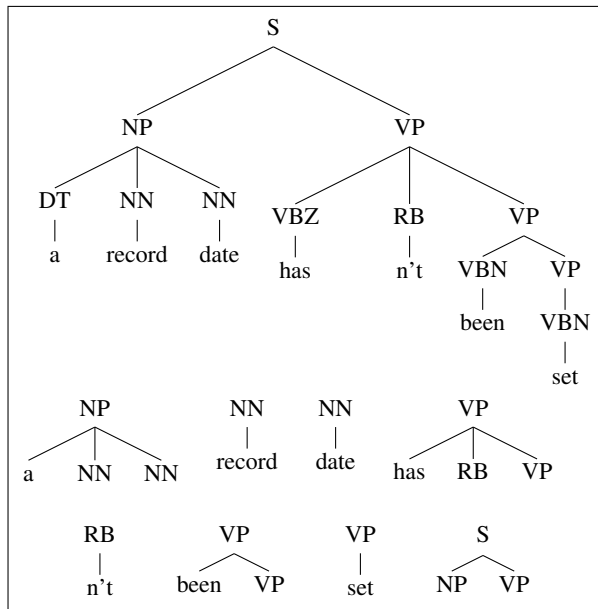


Figure 2: Example rules learned from the parse of the sentence “a record date has n’t been set”.

the same time returning low probability parses, if a parse at all, over examples of ungrammatical English.

We now briefly describe rule extraction. Given a sentence, we work in a left to right manner over the words. Taking a particular word, we find the nearest ancestor NT which has more than one child. We say that there is a path of NTs between the word and this ancestor. If there are no children to the left of this recently traversed path, then we extract a rule, with the ancestor as the LHS terminal, the word as the first item of the RHS, and then the remaining NT children of the ancestor are appended to the RHS. Figure 1 demonstrates this extraction process given a toy subtree.

If the ancestor does have children to the left of the recently traversed path, we then extract a rule with the child of the ancestor in the recently traversed path as the LHS NT, and the word is the RHS. Notice that these two rules are in the strict Greibach Normal Form.

As these rules are often insufficient to re-construct a tree, we extract all non-lexical rules that would be required to re-construct the training tree. It is the inclusion of these non-lexicalized rules which mean we interpret our grammar to be in a ‘loose’ Greibach Normal Form.

Figure 2 gives an example of the rules we learn from a parse over the sentence “a record date has n’t been set”. As can be seen from the examples, we lose information about the part-of-speech (POS) NTs attached to certain words in a parse, as well as any unit productions. However, as we

| Training Data | PCFG | LE | CM2 | 3-gram | 5-gram |
|---------------|---------|--------|---------------|--------|--------|
| PTB | 53.80 % | 60.16% | 67.07% | 53.50% | 53.71% |

Table 2: Percentage of references with higher probability than respective SMT output.

wish to use our grammar in an SMT system, we feel that the syntactic expectations of a lexical item are more important than the recovery of it’s correct POS.

3.5. Parsing Scores

In Table 1 we report Labelled Recall (LR), Labelled Precision (LP) and F1 scores for the different parsers, based on the PARSEVAL measures used in (Collins, 1997). The parsers were trained on sections 02-21 of the Wall Street Journal portion of the Penn Tree Bank (PTB) Marcus et al. (1994), about 40,000 sentences, and tested on section 23, about 2,500 sentences. A held-out data set was used to estimate probabilities over unseen words for our models. Probabilities were estimated in the standard manner using Maximum Likelihood, conditioning probabilities on the LHS of a rule.

The structure for parses produced by the Lexicalized Expectation model are slightly different to that of gold parses. For example, given the sub tree (NP (DT the) (NN car)), our model would learn the rules NP \rightarrow the NN and NN \rightarrow car. Given this grammar, asked to parse the sentence “the car”, we would return the sub tree (NP the (NN car)). This format is not accepted by the PARSEVAL script we use, and therefore alter the parse so it becomes (NP (DUMMY the) (NN car)). Because we lose information about the NT attached to certain lexical items, we see an expected drop in Tagging Accuracy and Labelled Recall (LR), however Labelled Precision (LP) is comparable to that of the basic PCFG parser.

4. Experiments

Formally, our experiment is composed of a set of SMT output and their respective reference translations, which in our case is 4 each. We say a sentence set is a specific SMT output sentence with its respective references. We define the classification accuracy of a model to be the percentage of references which are assigned an equal or higher probability than the MT output.

We translated 1797 Arabic sentences into English using a phrase-based SMT system similar to that described in (Koehn, 2004; Carter et al., 2008), to give us the disfluent sentences required. There are four reference candidates for each translation. The BLEU score for the mt08 test set is 37.99. For computational reasons, we removed SMT output sentences with a sentence length greater than 30, along with their references. This left us with 1078 sentences in total.

We show in Table 2 the percentage of reference sentences which were assigned a higher or equal probability to the respective SMT output by each of the different models.

Even though Och et al. (2004) report having unsuccessfully tried a unigram based length normalization technique, in our experiments all results are reported using simple length normalization, where the probability is divided by

the sentence length, as we found this to have a positive impact, as one would expect. Dividing the probability by the number of productions used did not work.

| # | BLEU | PCFG | LE | CM2 | 3-gram |
|------|---------|-------------|---------------|---------------|-------------|
| 1040 | 0 - 4 | 55.29% | 60.38% | 64.33% | 51.54% |
| 4 | 5 - 9 | 100% | 100% | 100% | 100% |
| 92 | 10 - 14 | 61.69% | 69.57% | 77.17% | 58.70% |
| 228 | 15 - 19 | 60.53% | 68.86% | 77.63% | 61.40% |
| 320 | 20 - 24 | 51.25% | 55.31% | 69.69% | 54.38% |
| 396 | 25 - 29 | 53.03% | 60.35% | 65.15% | 51.26% |
| 328 | 30 - 34 | 53.35% | 59.76% | 68.29% | 55.79% |
| 420 | 35 - 39 | 57.14% | 63.81% | 66.67% | 45.48% |
| 336 | 40 - 44 | 51.98% | 59.52% | 63.99% | 60.42% |
| 296 | 45 - 49 | 52.03% | 57.77% | 70.27% | 58.78% |
| 304 | 50 - 54 | 55.29% | 56.25% | 60.20% | 48.36% |
| 172 | 55 - 59 | 53.49% | 62.21% | 65.12% | 57.36% |
| 152 | 60 - 64 | 55.26% | 59.21% | 64.47% | 48.03% |
| 92 | 65 - 69 | 44.57% | 51.09% | 55.43% | 56.52% |
| 56 | 70 - 74 | 46.53% | 51.79% | 48.21% | 41.07% |
| 24 | 75 - 79 | 58.33% | 62.50% | 58.33% | 62.50% |
| 20 | 80 - 84 | 50% | 55% | 55% | 80% |
| 4 | 85 - 89 | 25% | 25% | 0% | 0% |
| 8 | 90 - 94 | 37.50% | 50% | 25% | 37.50% |
| 0 | 95 - 99 | n/a | n/a | n/a | n/a |
| 20 | 100 | 75% | 75% | 60% | 70% |

Table 3: Classification accuracy bucketed by sentence BLEU score of SMT output. # represents the number of reference sentences in each BLEU bucket.

| # | Lengths | PCFG | LE | CM2 | 3-gram |
|-----|---------|--------|---------------|---------------|--------|
| 177 | 1 - 4 | 49.72% | 51.98% | 57.63% | 47.76% |
| 340 | 5 - 9 | 44.12% | 47.94% | 48.53% | 43.82% |
| 416 | 10 - 14 | 52.40% | 55.77% | 55.77% | 48.56% |
| 455 | 15 - 19 | 51.87% | 55.82% | 64.40% | 50.53% |
| 438 | 20 - 24 | 50.23% | 56.39% | 62.10% | 48.63% |
| 468 | 25 - 29 | 54.27% | 61.11% | 70.73% | 54.70% |
| 492 | 30 - 34 | 49.80% | 60.77% | 65.24% | 52.03% |
| 464 | 35 - 39 | 50.65% | 56.90% | 69.18% | 51.94% |
| 370 | 40 - 44 | 56.49% | 64.32% | 71.89% | 58.65% |
| 334 | 45 - 49 | 60.78% | 70.06% | 82.04% | 67.37% |
| 201 | 50 - 54 | 73.63% | 81.59% | 88.06% | 69.15% |
| 92 | 55 - 59 | 67.39% | 70.65% | 84.78% | 54.35% |
| 44 | 60 - 64 | 79.55% | 88.64% | 93.18% | 68.18% |
| 18 | 65 - 69 | 83.33% | 77.78% | 88.89% | 83.33% |
| 3 | 70 - 74 | 66.67% | 100% | 100% | 33.33% |

Table 4: Classification accuracy bucketed by sentence length. # represents the number of reference sentences in each length bucket.

To see to what extent these findings hold across different BLEU scores, we bucketed the reference sentences by the BLEU score of the SMT output. Results are displayed in Table 3. Unsurprisingly, classification accuracy appears to decrease as BLEU increases; the better the SMT output, the harder it is for all models to differentiate between

| SMT sentences | Reference sentences |
|---|--|
| the spokesman added that four of afghan police officers who were within the group abducted , were freed . | the spokesman added that four afghan policemen who were among the kidnapped have been released . |
| some of the names of the press iraqi reveals the jordanians journalists who received aid from actors iraqi security . | iraqi journalist reveals names of jordanian journalists who received aid from iraqi security authorities . |
| the dream of a return to , and the crossing the homeland . | rafah crossing , and the dream of returning to homeland 's embrace . |
| an attempt to break the young and research in mental western the causes of the scientific excellence 2 | an attempt to penetrate the mentality of western youth , in search of causes of academic excellence 2 |

Table 5: Example SMT output and corresponding reference sentences, where the SMT sentence received a higher probability than the references by all models.

| Model | SMT | Ref1 | Ref2 | Ref3 | Ref4 |
|--------|---------------|-------|--------|--------|-------|
| PCFG | -9.91 | -9.41 | -6.96 | -7.4 | -7.6 |
| LE | -9.96 | -8.54 | -8.96 | -8.35 | -8.75 |
| CM2 | -10.92 | -7.82 | -10.13 | -10.22 | -8.39 |
| 3-gram | -2.73 | -2.69 | -2.71 | -2.76 | -2.76 |

Table 6: Average Log Probability assigned to the SMT output sentences, and each of the 4 corresponding reference sets. Figures in bold are the sets assigned the lowest probability by the models.

good and bad english. However the relative classification accuracy remains the same between the models.

To take a deeper look into which sentences the models were having problems with, we bucketed the reference sentences by their length, and examined the classification accuracy for each bucket for each model.

Results are displayed in Table 4. Surprisingly, CM2 has the highest classification accuracy for those reference sentences of length 1 – 4, outperforming the n-gram model. What is interesting is that the parsers actually appear to do better as the sentence length increases. Parsers traditionally perform worse on standard LR and LP measures as sentence length increases. This demonstrates that LR and LP based measures, which parsers have originally been developed and optimized towards, is not necessarily a good indicator of their ability to differentiate between good and bad English.

5. Discussion

Collins Model 2 outperforms the other parsers on this task, including the standard n-gram Language Models, suggesting that the use of an unaltered parser can outperform more basic PCFG’s and n-gram models when discriminating between good and bad English.

This is especially surprising as it was inferred that parsers would make poor LMs for SMT because they assigned a higher probability to the SMT output than the reference sentences. To see to what degree our findings differ from previously published results, we looked at the average log probability assigned to SMT output and reference sets by each of the parsers, as well as the trigram LM.

As can be seen in Table 6, all the models actually assign a higher average probability to the reference sets than the baseline SMT output. So it appears that parsers are able to discriminate between good and bad English, at least as

well as, if not better than, n-gram models trained on the same data.

6. Conclusions

In this paper, we have investigated the ability of parsers to differentiate between good and bad English. We have shown, in contrast to previous work, that parsers are actually able to differentiate between the two, at least when their probabilities are length normalized. Further we have shown how this improves as the sentences being compared get longer, implying an ability to help in resolving long range dependencies within SMT systems over and above n-gram language models.

References

- Bikel, Daniel M., 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of the second international conference on Human Language Technology Research*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Carter, Simon, Christof Monz, and Sirvan Yahyaei, 2008. The QMUL System Description for IWSLT 2008. In *Proc. of the International Workshop on Spoken Language Translation*. Hawaii, USA.
- Charniak, Eugene, Kevin Knight, and Kenji Yamada, 2003. Syntax-based language models for statistical machine translation. In *MT Summit IX. International Association for Machine Translation*.
- Collins, Michael, 1997. Three generative, lexicalized models for statistical parsing. In Philip R. Cohen and Wolfgang Wahlster (eds.), *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Somerset, New Jersey: Association for Computational Linguistics.
- Koehn, Philipp, 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas (AMTA)*.
- Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert Macintyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger, 1994. The Penn Treebank:

- Annotating predicate argument structure. In *ARPA Human Language Technology Workshop*.
- Och, Franz Josef, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev, 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the 2004 Meeting of the North American chapter of the Association for Computational Linguistics*.
- Post, Matt and Daniel Gildea, 2008. Parsers as language models for statistical machine translation. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA)*. Honolulu, HI.
- Stolcke, Andreas, 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–201.