GF Wide-coverage English-Finnish MT system for WMT 2015

Prasanth Kolachina
Computing Science
Chalmers University of Technology
prasanth.kolachina@cse.gu.se

Aarne Ranta
Computing Science
Chalmers University of Technology
aarne.ranta@cse.gu.se

Abstract

This paper describes the GF Wide-coverage MT system submitted to WMT 2015 for translation from English to Finnish. Our system uses a interlingua based approach, in which the interlingua is a shared formal representation, that abstracts syntactic structures over multiple languages. Our final submission is a re-ranked system in which we combine this baseline MT system with a factored LM model.

1 Introduction

Interlingual translation is an old idea that has been suggested numerous times and refuted almost as many times. A typical criticism is that the very idea is utopistic: that one can never build an interlingua that faithfully represents meaning in all languages of the world. However, as the focus in machine translation has shifted from the perfect rendering of meaning to less modest goals, the idea of an interlingua can be reconsidered.

In the current paper, we describe our system submission to the WMT shared task in the English-Finnish track. Our system is an interlingua-based system, the interlingua based on an *abstract syntax* in the sense of Grammatical Framework (GF) (Ranta, 2011). GF has been previously shown to work for domain-specific MT outperforming state-of-art systems using semantic interlinguas (Ranta et al., 2011). Departing from this, the GF wide-coverage Translator is an attempt following the current mainstream in the field of MT: we are content with browsing quality in the output of the MT systems, while achieving the low cost of interlingual MT systems. As such, the shared *abstract syntax* is mapped to different "surface" languages representing an abstraction of the deep syntactic structure for each of the languages. The abstraction from word order, morphology and certain deep syntactic phenomena, allows the interlingua to cope with unrelated languages. At the same time, these systems are scalable beyond toy examples, into wide-coverage systems.

We submit this system as our baseline over the English Finnish language pair for the WMT shared task. In addition, we also submitted a "re-ranked" variant of the same system as our primary submission, using statistical language models to re-score the translations from the baseline. Automatic evaluation metrics have shown small improvements from re-ranking our baseline system [1].

The paper is organized as follows: we describe our baseline system in Section 2 and the re-ranked variant in Section 3. We present our experiments and relevant discussion in Section 4.

2 GF Wide-coverage Translator

The GF Translator pipeline has three main phases:

- **Parsing** converts the source sentence into a forest of *abstract syntax trees* (AST), i.e. interlingual representations.
- **Disambiguation** selects the most probable AST.
- **Linearization** converts the AST into a sentence in each of the target languages.

Disambiguation is for efficiency reasons integrated in the parser, which enumerates the results lazily in order of decreasing probability (Angelov and Ljunglöf, 2014). Our current system performs disambiguation by using tree probabilities estimated from the Penn Treebank, converted into GF abstract syntax (Angelov, 2011). Unlike most K-best parsers, there is no upper limit on how many

[1] Scores obtained from http://matrix.statmt.org/

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 141–144,
Lisboa, Portugal, 17-18 September 2015.

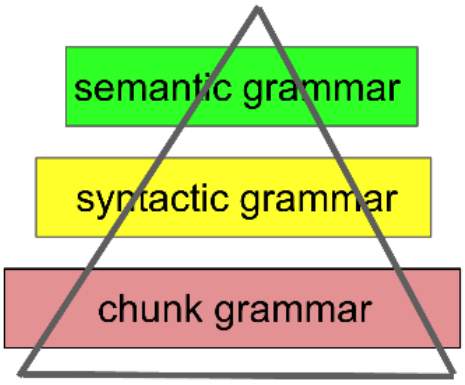
results can be obtained. Additionally, we use reversible mappings in our interlingua, thus reducing the work to define multilingual grammars for MT.

Translation is performed using the following components:

- A PGF grammar consisting of an abstract syntax (defining the ASTs) and, for each language, a concrete syntax that defines linearization and (by reversibility) parsing for the language.
- A probabilistic model for disambiguation
- The PGF interpreter, that consists of a generic parser and linearizer.

Since the PGF grammar forms a vital component of the MT system, we will now describe the wide-coverage grammar used in our system submission. All our submissions use this grammar as the "baseline". There is a large-scale single generic grammar based on the GF Resource Grammar Library (Ranta, 2009) that forms the central "backbone" of the wide-coverage grammar. As a whole, the grammar has the following components:

1. **RGL**, defining morphology and most of the syntax.
2. **Syntax extensions**, about 10% addition to RGL.
3. **Dictionary**, mapping abstract *word senses* to concrete words using open resources such as linked wordnets and wiktionaries (Virk et al., 2014); morphology mostly by the RGL's "smart paradigms" (Détrez and Ranta, 2012). Abstract dictionary entries are presented as English words split into distinct **senses**.
4. **Chunk grammar**, to make the translation robust for input that does not parse as complete sentences. It is inspired by Apertium(Forcada et al., 2011), which is a rule-based system operating only using chunks rather than deep syntactic analyses. In GF, it is derived from the RGL by enabling sub-sentential categories as start categories. The result can contain local agreement and reordering.
5. **Probabilities**, estimated from the Penn Treebank.



6. **CNL** using Semantic grammars, an optional part enabling domain adaptation via Embedded CNLs (Ranta, 2014). If something is parsable in the CNL, the CNL translation is given priority.

The GF Translator is not meant to be yet another browsing-quality system on the market. GF was originally designed for high-quality systems on specific domains. The novelty in our current system is that we can combine both coverage and quality in one and the same system. From the point of view of domain-specific applications, this means that the system does not just fail with out-of-grammar input as before, but offers robustness. From the open-domain point of view, the system offers a clear recipe for quality improvements by domain adaptation. In other words, the system we have built incorporates three levels of the Vauquois triangle in one and the same system: semantic, syntactic, and chunk-based translation, each of which and not just the highest level is based on its own part of the interlingua:

3 System Description

As mentioned in Section 1, our submission uses the GF Wide-coverage translator described in Section 2 as a baseline.

We are aware of one short-coming in the disambiguation model used in the baseline: the inference by the parser is carried out by context-free approximations. The context-free approximation is a reasonable approximation in the monolingual parsing scenario as shown by previous works in parsing literature. However, in the translation problem, the context-free assumption provides a poor approximation for inference. A simple example to illustrate this is the problem of sense selection by the parser. The choice of selecting a

particular word sense depends on both local contexts and entire sentential context. For e.g. the word "time" can refer to the sense that refers to temporality or the number of an attempt (as in *first time* or *hundredth time*). The choice of sense in this example can be made using surface context or *n-gram* information. Motivated primarily by this, we developed a re-ranked variant of the baseline system as described below.

Our re-ranked system re-estimates the scores of the K-best translations from the baseline using a linear mixture model. The mixture model uses the tree probability score obtained from the disambiguation model of the baseline system as the primary component. Each hypothesis in the K-best list is augmented using scores from n-gram language model (LM) that estimates the likelihood of the surface translations. Since our baseline system is an interlingua-based system, it is possible to integrate LM over multiple languages as different components in our mixture model. The resulting model selects the best translation by choosing the hypothesis with both the highest scoring abstract syntax tree and the best linearization of the abstract syntax tree.

4 Experiments

As part of the shared task contest, we carried out experiments with the wide-coverage translator and its re-ranked variant on the English-Finnish track. Table 1 shows the scores obtained by automatic evaluation for our system submissions.

On the *devel* set, the baseline system takes 27 minutes to carry out the translation pipeline i.e. the 1-best parsing of the English sentences combined with the 1-best linearization into Finnish. In comparison, the *test* set takes about 22 minutes for the pipeline. Of the 1500 sentences in the *devel* dataset, 600 sentences are parsed by the full RGL grammar, while the rest of the 900 sentences are parsed using the chunking grammar. We obtained similar statistics on the *test* dataset, where 560 sentences were parsed by the RGL and 810 sentences using the chunking grammar. This version of our translation pipeline is available online[2]. Manual evaluation and error analysis on a small sample from the *devel* dataset showed that the loss in MT quality from the chunking grammar was small, but significant. This is because the chunking grammar still allows for local agreement and reordering, while relaxing the RGL grammar. Nonetheless, we decided to use this version of the chunking grammar, without extending the RGL with new syntactic constructions. One reason for this decision was the speed up in the pipeline obtained by relaxing the full RGL grammar and adding the chunking grammar. It should be noted here that the quality of the MT system can be further improved by adding the full RGL at an additional computational cost. Evaluation experiments also showed that automatic evaluation metrics like BLEU substantially under-evaluate the perform of our system when used with a static translation as reference.

In the next round of experiments, we ran the parser and the linearizer in K-best modes, collecting the 50-best abstract syntax trees and the 30-best linearizations for each abstract syntax tree. Since the parsing and the linearization are carried out independent of one another, the 1500 hypothesis obtained from this run often contained identical translations. The overall number of distinct hypothesis in the K-best lists was typically found to be between 300 and 400. Collecting the K-best lists took about 93 minutes on the *devel* dataset and 80 minutes on the *test* dataset. We *reorder* these K-best lists using our reranking models, which consists of a re-scoring the hypothesis translations using a language model (LM) and estimating the mixed score for each hypothesis. The reordering combined with the re-scoring takes about 3-4 minutes on our lists of 1500-best hypotheses.

The LM for Finnish was trained on the Europarl corpus. Finnish sentences were morphologically analyzed and converted into a lemmatized corpus with morphological factors tagged along with the lemmas. We train a factored language model on this corpus, using the lemma and the part-of-speech and suffix as factors. In our current experiments, the hypothesis are re-scored using the Finnish language model alone, though in principle the re-scoring can be carried out using language models for multiple languages.

We train a ordinal regression model using the parse tree probability estimated using the GF disambiguation model and the factored LM score to re-order the K-best lists. A small set of 2500 sentences from the Europarl corpus were randomly taken and used as training samples for the regres-

[2] http://cloud.grammaticalframework.org/wc.html

System	BLEU	TER
Baseline	4.7	1.138
Reranked	4.8	1.135

Table 1: BLEU (11b) and TER scores obtained on the *newstest2015* dataset

sion model. The K-best lists in the training samples are ranked based on BLEU scores and TER scores.

Experiments with the *devel* dataset showed small improvements from using the LM to rescore the hypothesis. Comparatively, reranking resulted in even smaller improvements on the the *test* dataset. At this point, we carried out a analysis of the K-best lists on the *devel* set. We found that there was a very small variation in the K-best lists given the number of distinct hypothesis that were considered. Most of the variation was attributed to punctuation and orthography rather than *word senses* or *word order* as we initially expected.

Following this, we experimented with random sampling in the parse forests to evaluate the oracle quality of our translation system. The results of this study are pending error analysis and evaluation.

5 Conclusions

We described our system submission to the WMT shared task in the English-Finnish track in the current paper. Our system uses as interlingual-based approach, in which the interlingual is based on a shared representation of surface structures across languages. Our final submission is a hybrid system in which the K-best translations from the baseline system are re-ranked using a factored language model. We explain why our system results in a low-scoring baseline and discuss reasons why reranking provides minor improvements compared to previous approaches.

We plan to work on two extensions to the work described in this paper: first, we plan on increasing the variation in our K-best lists using sampling and incorporating heuristics into the parser. We hope that this will result in better improvements from re-ranking the K-best lists using a language model. Another extension we would like to experiment is the use of multiple language LMs to rescore the translations, this is uniquely possible only in our system since it allows for translation into multiple languages with little cost compared to other MT systems.

References

Krasimir Angelov and Peter Ljunglöf. 2014. Fast statistical parsing with parallel multiple context-free grammars. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 368–376, Gothenburg, Sweden, April. Association for Computational Linguistics.

Krasimir Angelov. 2011. *The Mechanics of the Grammatical Framework*. Ph.D. thesis, Chalmers University of Technology.

Grégoire Détrez and Aarne Ranta. 2012. Smart paradigms and the predictability and complexity of inflectional morphology. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–653, Avignon, France, April. Association for Computational Linguistics.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim ORegan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Aarne Ranta, Ramona Enache, and Grgoire Dtrez. 2011. Controlled Language for Everyday Use: the MOLTO Phrasebook. *Proceeding of CNL 2010, Zurich*.

A. Ranta. 2009. The GF Resource Grammar Library. *Linguistics in Language Technology*, 2. `http://elanguage.net/journals/index.php/lilt/article/viewFile/214/158`.

Aarne Ranta. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

Aarne Ranta. 2014. Embedded controlled languages. In *Controlled Natural Language - 4th International Workshop, CNL 2014, Galway, Ireland, August 20-22, 2014. Proceedings*.

Shafqat Mumtaz Virk, KVS Prasad, Aarne Ranta, and Krasimir Angelov. 2014. Developing an interlingual translation lexicon using wordnets and grammatical framework. *COLING 2014*, page 55.