STATISTICAL MACHINE TRANSLATION WITH LINGUISTIC LANGUAGE MODEL

Zuyev K. A. (konst@abbyy.com),
Indenbom E. M. (Eugene_I@abbyy.com),
Yudina M. V. (Maria_Yu@abbyy.com)

ABBYY, Moscow, Russia

Stemming from traditional "rule based" translation a "model based" approach is considered as an underlying model for statistical machine translation. This paper concerns with training on parallel corpora and application of this model for parsing and translation.

Preface

Statistical machine translation has made a significant breakthrough in machine translation within past decade. Due to availability of huge parallel corpora and increased raw computational power it turned out that rather simple statistical methods rival (and beat from commercial point of view) the traditional rule based methods with foundation on years of linguistic research. Nevertheless, the further advances in statistical machine translation are considered to be related with more linguistically-rich models. Even such a commodity tool as Moses provides support for using parsing information in translation process.

Statistical Machine Translation — a short overview

In statistical machine translation target sentences are produced from sentences by so-called "noisy-channel" — a filter, which modifies input into output. The design of true filter is unknown but can be modeled by assuming some parametric model. The model's parameters can be tuned and the structure can be validated by comparing behavior of model and "true filter". In case of machine translation the existing parallel corpora provide possible input and outputs for the modeled filter.

Originally models for statistical machine translation were very simple — a sequence of words. Then, to model the context dependency of translation, the phrase models and hierarchical phrase models were introduced [4]. It turned out that more complex models (with richer parametric space) are hard to trained. So parse trees are used to restrict possible phrases and labels familiar to linguists such as NP, VP are used to guess hierarchical phrases [6]. Actually now this model is a context free transduction grammar.

Although linguistic notions are used, little linguistic research is in place. Instead, the corpora marked-up with parse trees are used to train parsers.

Proposed approach

Language model used in our approach is based on well-established concepts of (noncomputational) linguistics. For the more detailed description, see [1]. Here is a brief summary of the model.

We represent a sentence by an HPSG-style tree. We distinguish between *surface* and *semantic structure*. Surface structure is language dependent, while semantic structure is deemed as universal.

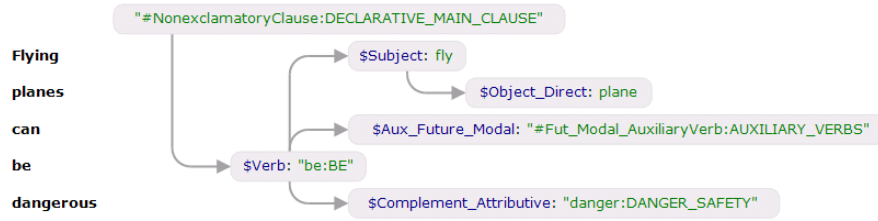
Therefore semantic structure is the "model" for translation process. Nodes of the tree (constituents) are normally formed from the words of the sentence. The constituent bears syntactic and semantic features. One of the most important features is *lexical class* — the representation of the meaning of the word. The meaning for our system is the position within our *semantic hierarchy*.

The *semantic hierarchy* (SH) — thesaurus-like hierarchical tree. It consists of universal nodes that represent different semantic concepts — semantic classes (SCs), which are filled with lexical items of natural languages — lexical classes (LCs). The main principle of organizing information within our hierarchy is the inheritance principle: higher nodes denote general notions, while their descendants denote more specific meaning and inherit main semantic and syntactic characteristics (these characteristics we call model) from their ancestors. Units of universal semantic information in our system are called *semantemes* — some of them are added in the hierarchy explicitly, others (for example, semantemes representing grammatical information such as tense, voice etc.) are computed during parsing.

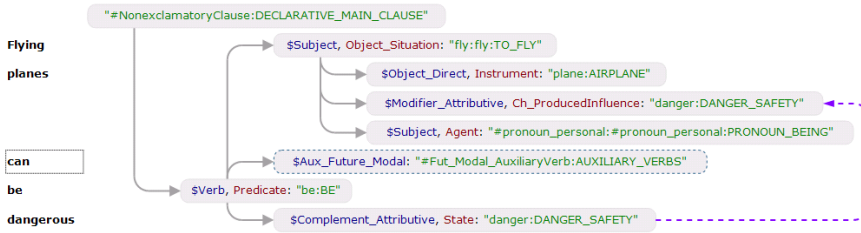
Dependencies between different units in the hierarchy are described in terms of semantic relations or *semantic slots* (which partly correlate to semantic roles, see [Fillmore 1968], for example). Semantic relations are also part of universal semantic structure and are language-independent. Dependencies between constituents on the surface syntactic level are called *surface slots* which are language-dependent. The correspondence between surface and semantic relation is called *diathesis*.

Along with tree dependencies, constituents can be linked with *non-tree relations* such as conjunction, anaphora, control and movement.

Syntactic structure



Semantic structure



For this model we have developed descriptions of semantic hierarchy, syntactic paradigms (surface slots with government, agreement, order restrictions and relations of slots and grammatical features. The descriptions distinguish between allowed and not allowed structures. There is no much emphasis on disambiguation of allowed structures.

Now we can reconsider translation process as conversion from source text to target via source surface structure, semantic structure, target surface structure to target text.

Since the model is ambiguous, we can treat this process as probabilistic and try to estimate conditional probabilities of the model features.

The probabilistic model includes:

- Lexeme & POS ngram probabilities
- Lexical class probability
- Lexicalized surface dependency probability
- Lexicalized semantic dependency probability
- Surface to sematic slot mapping
- Surface slots ngram probabilities
- Lexical classes co-occurrence probabilities
- Lexical class translation probability
- Surface slot translation probability

We use Bayesian approach to construct probability from different components. Taking into consideration the unprepared part of the audience of the conference we provide explanations instead of formulas.

Lexeme & POS ngram probabilities

This is a traditional language model, except that we take lexeme+part of speech instead of words. It is used to guide search on initial stages of parsing and in cases of incomplete parse trees. Currently we use 3-grams.

Lexical class probability

Lexical class probability differs significantly between various domains (e.g. meanings of word “file<noun>” in such domains as Law, Manufacturing or Information technologies). Thus, for the whole text we detect possible domains and calculate conditional

probability of different meaning of the words (lexical classes) for the Bayesian mixture of domains. For example, if we try to determine SC for the source lexeme *file* in the text for which we have established domain Information technologies, it is more realistic to choose the LC "file:FILE" (file as set of related data in computer). On the opposite, of we deal with the text labeled as Manufacturing domain, it is more probable that we have "file:FILE_AS_TOOL" ("a hand tool which is used for rubbing hard objects").

Processing of the whole text slightly improves precision of analysis and translation in comparison to sentence by sentence mode.

Lexicalized dependency probability

Lexicalized dependency probability (either surface or semantic) is a probability of the dependency link in the parse tree conditioned on lexical classes of parent and child. Currently there are ca. 500 dependency labels and more than 100K lexical classes. It means we have to learn more than $5 \times 1{,}012$ parameters.

Although many combinations are prohibited by the model, still their number is huge in comparison to the volume of available parallel corpora (~1G of words).

We use hierarchy to approximate parameters.

For example, if we try to determine the correct SC for *run* in the sentence like "I need to run the clock", we receive information from our hierarchy that *clock* is a device (the hole path up the tree is CLOCK: TIMEPIECE: DEVICE_FOR_MEASURING_AND_COUNTING: DEVICE), and we know that the class DEVICE is statistically good combined with the class " TO_ACTIVATE", so it is more reliable to choose "run:TO_ACTIVATE".

Lexicalized dependency probability is crucial for determination of the correct parsing tree and disambiguation of word senses.

Surface to sematic slot mapping

To select semantic slot for surface slot at analysis and to select surface slot for semantic slot at synthesis we collect co-occurrence data for surface and semantic slots.

Lexical classes co-occurrence probabilities

Domain depended lexical class probability provides only a rough adaptation to a particular large-scale domain. There are words, which senses do not correlate with easy identifiable domains or are indistinguishable within one, or there is no much text to identify domain and the dependency context is neutral. For example, in the sentence "Washington criticized Syria." we need to distinguish between the city and the surname (this difference does not influence translation, but is important for other applications of parsing). In this case co-occurrence of classes can help determine the right analysis if from the training data we know that Washington as a person had little to do with Syria.

Co-occurrence of classes is computed for siblings in dependency tree, for all words with limited neighborhood and for conjuncted words. Just as for the dependencies the number of parameters is quadratic to number of class. Here the approximation with hierarchy is used as well.

Lexical classes translation probability

Although the model was originally planned to have rich semantic features (*semantemes*) for differentiation between synonyms of one semantic class across languages, in practice we augmented it with conditional probability of synonym in target language for the give synonym in source language.

Surface slot translation probability

In theory, surface slot selection at target language must be guided by source semantic slot and features of child and parent constituents. But it turns out that it is not possible to take into account all cases in the model. Thus we use as well probabilistic model which estimates target diathesis probability by source surface slot and complexity of child subtree.

Hierarchical approximation of lexicalized pairwise correlations

Here we present our method of computing co-occurrence statistics in case of lack of data by using semantic hierarchy.

The co-occurrence we need to compute is conditional probability

$$\log \frac{P(A \cap B)}{P(A)P(B)},$$

where A and B are two lexical classes. In case we have enough data we can use counts to calculate this value

$$\log \frac{N(A \cap B)N}{N(A)N(B)},$$

But for many class pairs $N(A \cap B)$ is either very small (which makes very unreliable estimations) or zero. The required probability can be decomposed with the use of hierarchy: , where — is *i*th ancestor of A

$$\prod_{\substack{n=0\ldots L-1 \\ m=0\ldots K-1}} \frac{N(A^{(n)} \cap B^{(m)})N(A^{(n+1)} \cap B^{(m+1)})}{N(A^{(n)} \cap B^{(m+1)})N(A^{(n+1)} \cap B^{(m)})}, \text{ where } A^{(i)} \text{ — is } i^{\text{th}} \text{ ancestor of A}$$

Thus we can use counts of events for the classes in higher levels of hierarchy. These counts of superclasses are larger and give more accurate estimates of probability.

Training the probabilistic model

To train the model we have to have correct parse trees to estimate probabilities of model components. There is no such resource of adequate size. To cope with this problem we use parallel corpora and the parse trees are “hidden variables”.

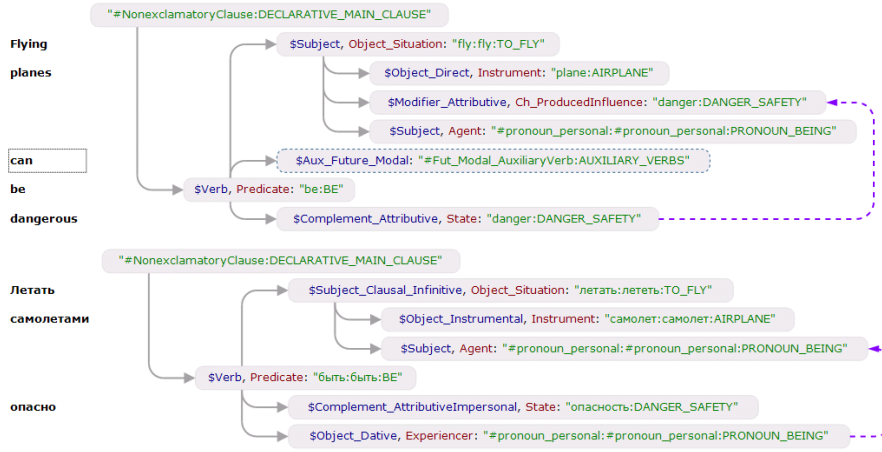
To make it work we need to have alignment of trees and a way to generate aligned parse trees. Alignment model is very simple — we condition the probability of alignment on, distance within hierarchy, on whether there are the same dependencies in aligned trees, and on the order of aligned constituents. To correctly handle lists of out-of-dictionary words (for example named entities) we also compute transliteration distance for such words.

To guess about hidden variable, that is presumably correct parse trees, we modified our parsing algorithm in the following way:

- We align two dependency graphs and attribute more weight to aligned constituents and links.
- We generate parse trees from the two graphs. They are generated by order of diminishing probability of parse structure to be correct and to produce the available translation.
- We align pairs of parse trees and select best trees (both by parsing and alignment quality).
- For further parameter estimation we utilize several generated trees to mitigate overfitting to erroneous parsing results.

See the example below on how the universal semantic structure and the parallel analysis help disambiguated classical case.

Parallel semantic structures



Resent research is concentrated on computing probabilistic model parameters for other linguistic descriptions such control, movement and ellipsis.

We also experimenting with non-Bayesian estimation of parameters, since Bayesian approach assumes independence of features which is hard to achieve.

Out of model translation

It is not feasible to cover complex, huge and dynamic languages by manual model. Two problems that we see are:

- There are too many words.
- Many contextual translations go across the hierarchy.

To cope with the first problem we have introduced a special lexeme for unknown words. We predict the morphological features of unknown (to our system) word by making hypothesis about its flexion. Unknown word lexeme is mapped to different places in the hierarchy, thus we also try to guess the rough meaning of the word, e.g. person, action, artifact.

At present, we either transliterate the unknown word or keep them untranslated. We could as well mine possible translation from alignments of parallel corpora.

The second problem is that some words in some context are translated in the adjacent or sometimes very far lexical classes of hierarchy (e.g. power plant — [электро] станция). As with phrase-based statistical machine translation we automatically capture regular out-of-hierarchy translations and use them as collocations. In comparison to phrases in SMT and collocations in traditional dictionaries our collocations are parse tree fragments. For more about mining the collocation, see [7].

To achieve the good quality of translation, comparable to popular online services, the system should be trained on huge, kept up-to-date internet corpora. Currently we train our system on roughly 10^8 sentences.

Evaluation

Internal evaluations

Internal evaluation is performed on several parallel and marked-up corpora.

We use modified BLEU to estimate translation quality. To our opinion, this variant of BLEU is more suitable for flective languages — only 1-gramms are matched literally, while higher-order n-gramms are reduced to lemmas. Absolute BLUE-score is very dependent to the corpora and to the system. For us it is 0.15-0.20. We rely on it to control incremental changes in the model and the algorithm.

Some corpora are partially marked-up with surface and semantic dependencies and lexical classes. We control the sentence level precision which is within 60-80%.

We also have small internal stand-out corpora to manually estimate and compare the translation quality with other systems.

External evaluations

It is hard to compare parsing performance of different systems if they are based on different linguistic principles. Anyway such attempt has been done at previous Dialog conferences. In [3] the part of speech disambiguation was tested (which indirectly correlates with parsing performance if parsing is used for this purpose). In [2] the parsing structures of different systems have been manually compared with a certain degree of freedom to match different approaches to the syntax. In both evaluations the system has shown good results.

This year the translation quality is estimated by range of automatic scores and by manual translation. We have achieved good results in both comparisons. The system was run in per-sentence mode without utilizing surrounding context. Although this context was available we were not able to use due to technical problems.

Conclusion

The development of the system and its good results in evaluations proves the plausibility of the linguistically oriented model-based approach to natural language processing. Due to the universality of the model, it can be used in many NLP tasks. Trained on the parallel corpora it can then perform translation, parsing, word sense disambiguation.

References

1. *Anisimovich, Druzhkin, Minlos, Petrova, Selegey, Zuev,* 2012. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. Proceedings of Dialog 2012 (pp. 80–103), Moscow, Russia.
2. *Toldova S. Ju.* et al. NLP evaluation 2011–2012: Russian syntactic parsers. Proceedings of Dialog 2012, Moscow, Russia.
3. *Lyashevskaya O., Astaf'eva I., Bonch-Osmolovskaya A., Garejshina A.,Grishina Ju., D'yachkov V., Ionov M., Koroleva A., Kudrinsky M., Lityagina A., Luchina E., Sidorova E., Toldova S., Savchuk S., Koval' S.* (2010), Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka [NLP evaluation: Russian morphological parsers], in Computational linguisticsand intellectual technologies. Proceedings of International Workshop Dialogue' 2010, Vol. 9 (16), Moscow, pp. 318–326.
4. *David Chiang.* 2005. A hierarchical phrase-based model for statistical machine translation. In Proceeding of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 263–270.
5. *Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst.* 2007. Moses: Open source toolkit for statistical machine translation. In 45th Annual Meeting of the Association for Computational Linguistics.
6. *Philipp Koehn, Barry Haddow, PhilipWilliams, and Hieu Hoang.* 2010. More linguistic annotation for statistical machine translation. In Proceedings of the Joint FifthWorkshop on Statistical Machine Translation and MetricsMATR, pages 115–120.
7. *Novitskiy, V. I.* Automatic retrieval of parallel collocations / V. I. Novitskiy // Pattern Recognition and Machine Intelligence / Ed. by S. Kuznetsov, D. Mandal, M. Kundu, S. Pal. — Vol. 6744 of Lecture Notes in Computer Science.— Moscow, Russia: Springer, 2011. — July. — pp. 261–267