Omnifluent^{TM} English-to-French and Russian-to-English Systems for the 2013 Workshop on Statistical Machine Translation

Evgeny Matusov, Gregor Leusch
Science Applications International Corporation (SAIC)
7990 Science Applications Ct.
Vienna, VA, USA
{evgeny.matusov,gregor.leusch}@saic.com

Abstract

This paper describes Omnifluent^{TM} Translate – a state-of-the-art hybrid MT system capable of high-quality, high-speed translations of text and speech. The system participated in the English-to-French and Russian-to-English WMT evaluation tasks with competitive results. The features which contributed the most to high translation quality were training data sub-sampling methods, document-specific models, as well as rule-based morphological normalization for Russian. The latter improved the baseline Russian-to-English BLEU score from 30.1 to 31.3% on a held-out test set.

1 Introduction

Omnifluent Translate is a comprehensive multilingual translation platform developed at SAIC that automatically translates both text and audio content. SAIC's technology leverages hybrid machine translation, combining features of both rule-based machine and statistical machine translation for improved consistency, fluency, and accuracy of translation output.

In the WMT 2013 evaluation campaign, we trained and tested the Omnifluent system on the English-to-French and Russian-to-English tasks. We chose the En–Fr task because Omnifluent En–Fr systems are already extensively used by SAIC's commercial customers: large human translation service providers, as well as a leading fashion designer company (Matusov, 2012). Our Russian-to-English system also produces high-quality translations and is currently used by a US federal government customer of SAIC.

Our experimental efforts focused mainly on the effective use of the provided parallel and monolingual data, document-level models, as well using rules to cope with the morphological complexity of the Russian language. While striving for the best possible translation quality, our goal was to avoid those steps in the translation pipeline which would make a real-time use of the Omnifluent system impossible. For example, we did not integrate re-scoring of N-best lists with huge computationally expensive models, nor did we perform system combination of different system variants. This allowed us to create a MT system that produced our primary evaluation submission with the translation speed of 18 words per second[1]. This submission had a BLEU score of 24.2% on the Russian-to-English task[2], and 27.3% on the English-to-French task. In contrast to many other submissions from university research groups, our evaluation system can be turned into a fully functional, commercially deployable on-line system with the same high level of translation quality and speed within a single work day.

The rest of the paper is organized as follows. In the next section, we describe the core capabilities of the Omnifluent Translate systems. Section 3 explains our data selection and filtering strategy. In Section 4 we present the document-level translation and language models. Section 5 describes morphological transformations of Russian. In sections 6 we present an extension to the system that allows for automatic spelling correction. In Section 7, we discuss the experiments and their evaluation. Finally, we conclude the paper in Section 8.

2 Core System Capabilities

The Omnifluent system is a state-of-the-art hybrid MT system that originates from the AppTek technology acquired by SAIC (Matusov and Köprü, 2010a). The core of the system is a statistical search that employs a combination of multiple

[1] Using a single core of a 2.8 GHz Intel Xeon CPU.

[2] The highest score obtained in the evaluation was 25.9%

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 158–163,
Sofia, Bulgaria, August 8-9, 2013 ©2013 Association for Computational Linguistics

probabilistic translation models, including phrase-based and word-based lexicons, as well as reordering models and target n-gram language models. The retrieval of matching phrase pairs given an input sentence is done efficiently using an algorithm based on the work of (Zens, 2008). The main search algorithm is the source cardinality-synchronous search. The goal of the search is to find the most probable segmentation of the source sentence into non-empty non-overlapping contiguous blocks, select the most probable permutation of those blocks, and choose the best phrasal translations for each of the blocks at the same time. The concatenation of the translations of the permuted blocks yields a translation of the whole sentence. In practice, the permutations are limited to allow for a maximum of M "gaps" (contiguous regions of uncovered word positions) at any time during the translation process. We set M to 2 for the English-to-French translation to model the most frequent type of reordering which is the reordering of an adjective-noun group. The value of M for the Russian-to-English translation is 3.

The main differences of Omnifluent Translate as compared to the open-source MT system Moses (Koehn et al., 2007) is a reordering model that penalizes each deviation from monotonic translation instead of assigning costs proportional to the jump distance (4 features as described by Matusov and Köprü (2010b)) and a lexicalization of this model when such deviations depend on words or part-of-speech (POS) tags of the last covered and current word (2 features, see (Matusov and Köprü, 2010a)). Also, the whole input document is always visible to the system, which allows the use of document-specific translation and language models. In translation, multiple phrase tables can be interpolated linearly on the count level, as the phrasal probabilities are computed on-the-fly. Finally, various novel phrase-level features have been implemented, including binary topic/genre/phrase type indicators and translation memory match features (Matusov, 2012).

The Omnifluent system also allows for partial or full rule-based translations. Specific source language entities can be identified prior to the search, and rule-based translations of these entities can be either forced to be chosen by the MT system, or can compete with phrase translation candidates from the phrase translation model. In both cases, the language model context at the boundaries of the rule-based translations is taken into account. Omnifluent Translate identifies numbers, dates, URLs, e-mail addresses, smileys, etc. with manually crafted regular expressions and uses rules to convert them to the appropriate target language form. In addition, it is possible to add manual translation rules to the statistical phrase table of the system.

3 Training Data Selection and Filtering

We participated in the constrained data track of the evaluation in order to obtain results which are comparable to the majority of the other submissions. This means that we trained our systems only on the provided parallel and monolingual data.

3.1 TrueCasing

Instead of using a separate truecasing module, we apply an algorithm for finding the true case of the first word of each sentence in the target training data and train truecased phrase tables and a truecased language model[3]. Thus, the MT search decides on the right case of a word when ambiguities exist. Also, the Omnifluent Translate system has an optional feature to transfer the case of an input source word to the word in the translation output to which it is aligned. Although this approach is not always error-free, there is an advantage to it when the input contains previously unseen named entities which use common words that have to be capitalized. We used this feature for our English-to-French submission only.

3.2 Monolingual Data

For the French language model, we trained separate 5-gram models on the two GigaWord corpora AFP and APW, on the provided StatMT data for 2007–2012 (3 models), on the EuroParl data, and on the French side of the bilingual data. LMs were estimated and pruned using the IRSTLM toolkit (Federico et al., 2008). We then tuned a linear combination of these seven individual parts to optimum perplexity on WMT test sets 2009 and 2010 and converted them for use with the KenLM library (Heafield, 2011). Similarly, our English LM was a linear combination of separate LMs built for GigaWord AFP, APW, NYT, and the other parts, StatMT 2007–2012, Europarl/News Commentary, and the Yandex data, which was tuned for best perplexity on the WMT 2010-2013 test sets.

[3] Source sentences were lowercased.

3.3 Parallel Data

Since the provided parallel corpora had different levels of noise and quality of sentence alignment, we followed a two-step procedure for filtering the data. First, we trained a baseline system on the "good-quality" data (Europarl and News Commentary corpora) and used it to translate the French side of the Common Crawl data into English. Then, we computed the position-independent word error rate (PER) between the automatic translation and the target side on the segment level and only kept those original segment pairs, the PER for which was between 10% and 60%. With this criterion, we kept 48% of the original 3.2M sentence pairs of the common-crawl data.

To leverage the significantly larger Multi-UN parallel corpus, we performed perplexity-based data sub-sampling, similarly to the method described e. g. by Axelrod et al. (2011). First, we trained a relatively small 4-gram LM on the source (English) side of our development data and evaluation data. Then, we used this model to compute the perplexity of each Multi-UN source segment. We kept the 700K segments with the lowest perplexity (normalized by the segment length), so that the size of the Multi-UN corpus does not exceed 30% of the total parallel corpus size. This procedure is the only part of the translation pipeline for which we currently do not have a real-time solution. Yet such a real-time algorithm can be implemented without problems: we word-align the original corpora using GIZA++ahead of time, so that after sub-sampling we only need to perform a quick phrase extraction. To obtain additional data for the document-level models only (see Section 4), we also applied this procedure to the even larger Gigaword corpus and thus selected 1M sentence pairs from this corpus.

We used the PER-based procedure as described above to filter the Russian-English Common-crawl corpus to 47% of its original size. The baseline system used to obtain automatic translation for the PER-based filtering was trained on News Commentary, Yandex, and Wiki headlines data.

4 Document-level Models

As mentioned in the introduction, the Omnifluent system loads a whole source document at once. Thus, it is possible to leverage document context by using document-level models which score the phrasal translations of sentences from a specific document only and are unloaded after processing of this document.

To train a document-level model for a specific document from the development, test, or evaluation data, we automatically extract those source sentences from the background parallel training data which have (many) n-grams (n=2...7) in common with the source sentences of the document. Then, to train the document-level LM we take the target language counterparts of the extracted sentences and train a standard 3-gram LM on them. To train the document-level phrase table, we take the corresponding word alignments for the extracted source sentences and their target counterparts, and extract the phrase table as usual. To keep the additional computational overhead minimal yet have enough data for model estimation, we set the parameters of the n-gram matching in such a way that the number of sentences extracted for document-level training is around 20K for document-level phrase tables and 100K for document-level LMs.

In the search, the counts from the document-level phrase table are linearly combined with the counts from the background phrase table trained on the whole training data. The document-level LM is combined log-linearly with the general LM and all the other models and features. The scaling factors for the document-level LMs and phrase tables are not document-specific; neither is the linear interpolation factor for a document-level phrase table which we tuned manually on a development set. The scaling factor for the document-level LM was optimized together with the other scaling factors using Minimum Error Rate Training (MERT, see (Och, 2003)).

For English-to-French translation, we used both document-level phrase tables and document-level LMs; the background data for them contained the sub-sampled Gigaword corpus (see Section 3.3). We used only the document-level LMs for the Russian-to-English translation. They were extracted from the same data that was used to train the background phrase table.

5 Morphological Transformations of Russian

Russian is a morphologically rich language. Even for large vocabulary MT systems this leads to data sparseness and high out-of-vocabulary rate. To mitigate this problem, we developed rules for reducing the morphological complexity of the language, making it closer to English in terms of the used word forms. Another goal was to ease the translation of some morphological and syntactic phenomena in Russian by simplifying them; this included adding artificial function words.

We used the *pymorphy* morphological analyzer[4] to analyze Russian words in the input text. The output of *pymorphy* is one or more alternative analyses for each word, each of which includes the POS tag plus morphological categories such as gender, tense, etc. The analyses are generated based on a manual dictionary, do not depend on the context, and are not ordered by probability of any kind. However, to make some functional modifications to the input sentences, we applied the tool not to the vocabulary, but to the actual input text; thus, in some cases, we introduced a context dependency. To deterministically select one of the *pymorphy*'s analyses, we defined a POS priority list. Nouns had a higher priority than adjectives, and adjectives higher priority than verbs. Otherwise we relied on the first analysis for each POS.

The main idea behind our hand-crafted rules was to normalize any ending/suffix which does not carry information necessary for correct translation into English. Under normalization we mean the restoration of some "base" form. The *pymorphy* analyzer API provides inflection functions so that each word could be changed into a particular form (case, tense, etc.). We came up with the following normalization rules:

- convert all adjectives and participles to first-person masculine singular, nominative case;
- convert all nouns to the nominative case keeping the plural/singular distinction;
- for nouns in genitive case, add the artificial function word "of_" after the last noun before the current one, if the last noun is not more than 4 positions away;
- for each verb infinitive, add the artificial function word "to_" in front of it;
- convert all present-tense verbs to their infinitive form;
- convert all past-tense verbs to their past-tense first-person masculine singular form;
- convert all future-tense verbs to the artificial function word "will_" + the infinitive;
- For verbs ending with reflexive suffixes ся/сь, add the artificial function word "sya_" in front of the verb and remove the suffix. This is done to model the reflexion (e.g. "он умывался"– "он sya_ умывал" – "he washed himself", here "sya_" corresonds to "himself"), as well as, in other cases, the passive mood (e.g. "он вставляется" – "он sya_ вставлять"– "it is inserted").

An example that is characteristic of all these modifications is given in Figure 1.

It is worth noting that not all of these transformations are error-free because the analysis is also not always error-free. Also, sometimes there is information loss (as in case of the instrumental noun case, for example, which we currently drop instead of finding the right artificial preposition to express it). Nevertheless, our experiments show that this is a successful morphological normalization strategy for a statistical MT system.

6 Automatic Spelling Correction

Machine translation input texts, even if prepared for evaluations such as WMT, still contain spelling errors, which lead to serious translation errors. We extended the Omnifluent system by a spelling correction module based on Hunspell[5] – an open-source spelling correction software and dictionaries. For each input word that is unknown *both* to the Omnifluent MT system and to Hunspell, we add those Hunspell's spelling correction suggestions to the input which are in the vocabulary of the MT system. They are encoded in a lattice and assigned weights. The weight of a suggestion is inversely proportional to its rank in the Hunspell's list (the first suggestions are considered to be more probable) and proportional to the unigram probability of the word(s) in the suggestion. To avoid errors related to unknown names, we do not apply spelling correction to words which begin with an uppercase letter.

The lattice is translated by the decoder using the method described in (Matusov et al., 2008); the globally optimal suggestion is selected in the translation process. On the English-to-French task, 77 out of 3000 evaluation data sentences were translated differently because of automatic spelling correction. The BLEU score on these sentences improved from 22.4 to 22.6%. Manual analysis of the results shows that in around

[4] `https://bitbucket.org/kmike/pymorphy`

[5] `http://hunspell.sourceforge.net`

source	Обед проводился в отеле Вашингтон спустя несколько часов после совещания суда по делу
prep	Обед sya_ проводил в отель Вашингтон спустя несколько часы после совещание of_ суд по дело
ref	The dinner was held at a Washington hotel a few hours after the conference of the court over the case

Figure 1: Example of the proposed morphological normalization rules and insertion of artificial function words for Russian.

System	BLEU	PER
	[%]	[%]
baseline	31.3	41.1
+ extended features	31.7	41.0
+ alignment combination	32.1	40.6
+ doc-level models	32.7	39.3
+ common-crawl/UN data	33.0	39.9

Table 1: English-to-French translation results (newstest-2012-part2 progress test set).

System	BLEU	PER
	[%]	[%]
baseline (full forms)	30.1	38.9
morph. reduction	31.3	38.1
+ extended features	32.4	37.3
+ doc-level LMs	32.3	37.4
+ common-crawl data	32.9	37.1

Table 2: Russian-to-English translation results (newstest-2012-part2 progress test set).

70% of the cases the MT system picks the right or almost right correction. We applied automatic spelling correction also to the Russian-to-English evaluation submissions. Here, the spelling correction was applied to words which remained out-of-vocabulary after applying the morphological normalization rules.

7 Experiments

7.1 Development Data and Evaluation Criteria

For our experiments, we divided the 3000-sentence newstest-2012 test set from the WMT 2012 evaluation in two roughly equal parts, respecting document boundaries. The first part we used as a tuning set for N-best list MERT optimization (Och, 2003). We used the second part as a test set to measure progress; the results on it are reported below. We computed case-insensitive BLEU score (Papineni et al., 2002) for optimization and evaluation. Only one reference translation was available.

7.2 English-to-French System

The baseline system for the English-to-French translation direction was trained on Europarl and News Commentary corpora. The word alignment was obtained by training HMM and IBM Model 3 alignment models and combining their two directions using the "grow-diag-final" heuristic (Koehn, 2004). The first line in Table 1 shows the result for this system when we only use the standard features (phrase translation and word lexicon costs in both directions, the base reordering features as described in (Matusov and Köprü, 2010b) and the 5-gram target LM). When we also optimize the scaling factors for extended features, including the word-based and POS-based lexicalized reordering models described in (Matusov and Köprü, 2010a), we improve the BLEU score by 0.4% absolute. Extracting phrase pairs from three different, equally weighted alignment heuristics improves the score by another 0.3%. The next big improvement comes from using document-level language models and phrase tables, which include Gigaword data. Especially the PER decreases significantly, which indicates that the document-level models help, in most cases, to select the right word translations. Another significant improvement comes from adding parts of the Common-crawl and Multi-UN data, sub-sampled with the perplexity-based method as described in Section 3.3. The settings corresponding to the last line of Table 1 were used to produce the Omnifluent primary submission, which resulted in a BLEU score of 27.3 on the WMT 2013 test set.

After the deadline for submission, we discovered a bug in the extraction of the phrase table which had reduced the positive impact of the extended phrase-level features. We re-ran the optimization on our tuning set and obtained a BLEU score of 27.7% on the WMT 2013 evaluation set.

7.3 Russian-to-English System

The first experiment with the Russian-to-English system was to show the positive effect of the morphological transformations described in Section 5. Table 2 shows the result of the baseline system, trained using full forms of the Russian words on the News Commentary, truecased Yandex and Wiki Headlines data. When applying the morphological transformations described in Section 5 both in training and translation, we obtain a significant improvement in BLEU of 1.3% absolute. The out-of-vocabulary rate was reduced from 0.9 to 0.5%. This shows that the morphological reduction actually helps to alleviate the data sparseness problem and translate structurally complex constructs in Russian.

Significant improvements are obtained for Ru–En through the use of extended features, including the lexicalized and "POS"-based reordering models. As the "POS" tags for the Russian words we used the *pymorphy* POS tag selected deterministically based on our priority list, together with the codes for additional morphological features such as tense, case, and gender. In contrast to the En–Fr task, document-level models did not help here, most probably because we used only LMs and only trained on sub-sampled data that was already part of the background phrase table. The last boost in translation quality was obtained by adding those segments of the cleaned Common-crawl data to the phrase table training which are similar to the development and evaluation data in terms of LM perplexity. The BLEU score in the last line of Table 2 corresponds to Omnifluent's BLEU score of 24.2% on the WMT 2013 evaluation data. This is only 1.7% less than the score of the best BLEU-ranked system in the evaluation.

8 Summary and Future Work

In this paper we described the Omnifluent hybrid MT system and its use for the English-to-French and Russian-to-English WMT tasks. We showed that it is important for good translation quality to perform careful data filtering and selection, as well as use document-specific phrase tables and LMs. We also proposed and evaluated rule-based morphological normalizations for Russian. They significantly improved the Russian-to-English translation quality. In contrast to some evaluation participants, the presented high-quality system is fast and can be quickly turned into a real-time system. In the future, we intend to improve the rule-based component of the system, allowing users to add and delete translation rules on-the-fly.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation via Pseudo In-Domain Data Selection. In *International Conference on Emperical Methods in Natural Language Processing*, Edinburgh, UK, July.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Proceedings of Interspeech*, pages 1618–1621.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

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 *Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *6th Conference of the Association for Machine Translation in the Americas (AMTA 04)*, pages 115–124, Washington DC, September/October.

Evgeny Matusov and Selçuk Köprü. 2010a. AppTek's APT Machine Translation System for IWSLT 2010. In *Proc. of the International Workshop on Spoken Language Translation*, Paris, France, December.

Evgeny Matusov and Selçuk Köprü. 2010b. Improving Reordering in Statistical Machine Translation from Farsi. In *AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado, USA, November.

Evgeny Matusov, Björn Hoffmeister, and Hermann Ney. 2008. ASR word lattice translation with exhaustive reordering is possible. In *Interspeech*, pages 2342–2345, Brisbane, Australia, September.

Evgeny Matusov. 2012. Incremental Re-training of a Hybrid English-French MT System with Customer Translation Memory Data. In *10th Conference of the Association for Machine Translation in the Americas (AMTA 12)*, San Diego, CA, USA, October-November.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Richard Zens. 2008. *Phrase-based Statistical Machine Translation: Models, Search, Training*. Ph.D. thesis, RWTH Aachen University, Aachen, Germany, February.