Large Scale Speech-to-Text Translation with Out-of-Domain Corpora using Better Context-Based Models and Domain Adaptation

Marcin Junczys-Dowmunt[1], *Paweł Przybysz*[1], *Arleta Staszuk*[1], *Eun-Kyoung Kim*[2], *JaeWon Lee*[2]

[1]Samsung Poland R&D Center, Warsaw, Poland
[2]DMC R&D Center, Samsung Electronics, Suwon, South Korea

m.junczys@samsung.com, p.przybysz@samsung.com, a.staszuk@samsung.com,
ekkim.kim@samsung.com, jwonlee@samsung.com

Abstract

In this paper, we described the process of building a large-scale speech-to-text pipeline. Two target domains, daily conversations and travel-related conversations between two agents, for the English-German language pair (both directions) are examined. The SMT component is built from out-of-domain but freely-available bilingual and monolingual data. We make use of most of the known available resources to examine the effects of unrestricted data and large scale models. A naive baseline delivers solid results in terms of MT-quality. Extending the baseline with context-based translation model features like operations sequence models, higher-order class-based language models, and additional web-scale word-based language models leads to a system that significantly outperforms the baseline. Domain adaption is performed by separately weighting the influence of the out-of-domain subcorpora. This is explored for translation models and language models yielding significant improvements in both cases. Automatic and manual evaluation results are provided for raw MT-quality and ASR+MT-quality.

Index Terms: Statistical Machine Translation, Speech-to-Text Translation, Large-scale SMT.

1. Introduction

Speech-to-Text MT in mobile settings has recently seen an increase in interest. We report the results of a feasibility experiment in Speech-to-Text MT involving mobile and smart devices while focusing on the SMT component of the system. SMT competitions like the WMT or IWSLT workshops restrict the permitted training data to specified data sets and try to focus on methods and algorithms alone. In contrast, we provide a description of a system that applies conclusions from recent WMT submissions to an unrestricted data setting and shows how useful these are to real-world applications outside the tight confinements of said shared tasks.

The paper is organized as follows: firstly we describe the usage scenario, the full pipeline architecture, and provide a short description of the in-house ASR component. A section with a full list of translation model and language model data follows. Next, a baseline SMT system is built from this data and improvements are added in the following section. An extensive evaluation section contains automatically and manually calculated evaluation results, for both, raw text-based MT-quality and the full ASR-MT-pipeline.

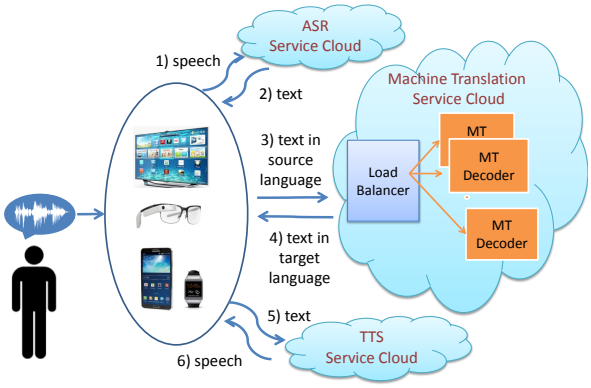

Figure 1: Pipeline Overview

2. General Scenario

Figure 1 illustrates a simplified representation of the intended application scenario. Currently, all systems pass on 1-best results of the respective component. Depending on the device used, MT-output can be displayed as text and/or passed on to a Text-to-Speech system. In this paper we omit the Text-to-Speech component and focus mainly on the MT module. The Speech-to-text component is treated as a black-box and not modified nor adapted to the specific task.

3. Training and Test Data

3.1. Training Data

The translation model has been built using most of the freely available parallel corpora for the English-German pair we know of, see Table 1a for a complete list of resources and their contribution to the model. The DGT-TM and JRC-Acquis 2.2 probably overlap in large parts which has not been further examined. With JRC-Acquies 3.0 available (which has not been used) it makes sense to retain only one of these resources in the future.

Language model data (Table 1b) for the later described baseline system incorporates all the respective target language data from the bilingual resources. For the improved systems, parts of the CommonCrawl-extracted data made available by [7] has been used. For the English language model we only processed the first ten (out of one hundred) shards of available text due to hardware and time restrictions. The available German

Table 1: Resources used for training

(a) Bilingual training data

Corpus name and source	No. of Sentences
Common Crawl Parallel Corpus [1]	2 399 123
The EU bookshop corpus [2]	9 890 525
Europarl v7 [3]	1 920 209
EMEA [2]	1 108 752
DGT-TM 2014 [4]	3 707 302
JRC-Acquis 2.2 [5]	1 211 220
OpenSubtitles2012 [2]	5 633 148
OpenSubtitles2013 [2]	5 269 507
PatTR [6]	22 998 357
Proprietary Bilingual Dictionary	1 765 814
Total	55 903 957

(b) Monolingual training data

Corpus name and source	No. of Sentences
All of the above for EN→DE and DE→EN	55 903 957
DE→EN: first 10% of the English CommonCrawl data [7]	ca. 5×10^{10}
EN→DE: All of the German CommonCrawl data [7]	ca. 3×10^{10}

data has been used in full. The raw text copora are therefore within the same order of magnitude for both languages.

None of the training data subcorpora can be classified as in-domain. There may however be matching fragments in some of them. We can for instance assume, that the CommonCrawl-based resources contain tourism-related data, while movie subtitles are naturally rich with dialogues.

3.2. Test and Development Data

In-domain test data has been collected in the following steps:

- Roughly 1500 English sentences and 1500 German sentences from conversational settings have been collected from various (non-overlapping) real world sources for each of the two domains.
- All sentences have been translated in their respective target language, which resulted in 3000 different sentence pairs for each domain.
- In order to create spoken language test sets, dialogues in both languages have been recorded by native speakers. Different roles in dialogues where filled by two different speakers.

Development data consists of 3000 randomly selected sentence pairs from TED talks [2], TED talks where fully excluded from the TM and LM training data. While this is not dialogue data is seems to fulfill its purpose: results do not differ significantly from cross tuning with the two target domains (using daily-dialogue as development set for travel-dialogue and vice versa).

4. Speech-to-Text

The used ASR systems are general purpose systems developed in-house. Both systems are still in development, the English system being in a much earlier state than its German counterpart. Results should be seen as tentative. Table 2 shows the performance on the used test sets. Similar results for various freely accessible English speech-to-text systems were described by [8] for a dialogue system. Our German system already holds up well, the English system falls off considerably. This will be improved in the future. The quality of the ASR system does of course affect the over-all quality of the ASR+MT system, as seen in the evaluation section.

Table 2: ASR performance

(a) German test sets

Word acc.	Sent. acc.
65.4	29.2

(b) English test sets

Word acc.	Sent. acc.
51.30	12.20

5. A WMT-inspired SMT System

5.1. Baseline

The baseline SMT system consists of a vanilla Moses [9] configuration. All training data is concatenated and treated as a single training corpus. Phrase-table extraction uses Good-Turing smoothing, the final phrase-table has been significance pruned [10] for size reduction.[1] We use the compact phrase table and reordering model representation for binarization [11].

The 5-gram language model are estimated with Modified Kneser-Ney smoothing [12, 13] from the respective target language data of the parallel corpus. To reduce size requirements, we use heavily quantized binary models with no noticeable quality reduction. Pruning is applied to all singleton n-grams with n equal to or greater than 3.

All results in this paper are reported after parameter tuning, adding new features resulted in repeated tuning.

5.2. Improving the Baseline with more Context

In this section we extend the baseline model with two additional context-based models: an Operation Sequence Model (OSM) that adds Markov-context to translation models, and a high-order class-based language model.

OSMs can be seen as bilingual language models that add context-dependency between minimal translation units in the otherwise independent phrases of the translation model. Other than the typical language model, probabilities are also modeled on the source side of the translation process. In a restricted data setting, [14] report on improvements for various language pairs when an OSM is added to the phrase-based decoder. We can confirm that this positive effect scales up to our large-data settings, see Table 3 for results.

Both target domains consist of mainly short sentences, by adding a higher-order language model we can capture a larger part or the whole target sentence. Class-based language models seem to be a good compromise between increased n-gram

[1] In our experiments significance pruning resulted in no quality loss while reducing translation model size by a factor of 5 and considerable speed improvements.

Table 3: Improvement of MT-Component in BLEU %

(a) German-to-English

System	Daily	Travel
Baseline	38.4	40.2
+OSM	38.8	40.4
+WCLM	39.1	40.6
+CC LM/WCLM	41.4	42.5
+TM-Domains	42.5	43.2
+LM-Domains	43.2	43.7

(b) English-to-German

System	Daily	Travel
Baseline	30.8	32.6
+OSM	31.0	33.5
+WCLM	31.3	33.6
+CC LM/WCLM	31.9	35.2
+TM-Domains	32.8	35.3
+LM-Domains	33.1	36.0

length and total model size. We use automatically calculated word cluster ids as classes. Brown-clusters [15] are a popular way to create automatic classes, but computation is expensive for larger data. Instead, we make use of `word2vec` [16] which can handle large corpus sizes and compute 200 word classes based on the target language data. Next, the target language corpus is mapped to sequences of class ids and a 9-gram language model is estimated. Again we see small but consistent improvements due to adding the word-class language models (denoted as WCLM) for both translation directions and domains.

5.3. Better Language Models

Larger language models estimated from the CommonCrawl data described in section 3.1 replace the previous models. Due to the enormous corpus sizes we need to apply more aggressive pruning, beginning with 3-grams we increase the pruning limit with n-grams order, i.e 3-grams need to occur at least two times, 4-grams three times, etc.

The scalability of `word2vec` is of advantage. We recompute word classes from the CommonCrawl corpora and create a new word class 9-gram language model, replacing the previous WCLM which was based on the target part of the parallel training data only. Again, an aggressive pruning scheme with increasing pruning limits for higher orders is used.

Replacing the previous language models with the CommonCrawl-based models (denote as CC LM and CC WCLM), we see very significant improvements for all domains and translation directions.

5.4. TM Domain Adaptation

Strictly speaking, none of our parallel training data is in-domain data in respect to our target domains. However, some subcorpora may be better suited for the required domains while others should be penalized. This is tested by adding domain indicators to the translation model. Each phrase pair is annotated with a 0-1 vector of length N, where N is the number of domains. If a phrase has been seen in a domain the corresponding value is set to 1 or to 0 otherwise. Weights for these vectors are learned during parameter optimization. Subcorpora names are used as domains, however, other domain types, e.g. automatically computed clusters, could be used. Since there are 10 different sources of parallel sentences in our training data, each phrase pair is annotated with 10 additional scores.

As it turns out, the standard optimization algorithm MERT [17] cannot cope with as many features (the model now has 33 dense features) and we tested other optimization methods with domain indicators. The best and most stable results have been achieved with PRO [18]. Inspecting the resulting weights, it is not surprising to see that, for instance, features corresponding to patent data receive negative weights while movie subtitles have positive weights.

5.5. LM Domain Adaptation

Following a similar idea as above for language modeling, we extend our configuration with additional domain-specific language models. Here, text from each domain is being turned into a separate language model and all models are combined in a log-linear combination as new features, the model weights can still be tuned separately. This adds another 10 language models to the configuration and results in 43 dense features in total. As before, we now resort to PRO for parameter tuning. This again leads to a significant improvement in translation quality.

Future work in domain adaptation in this scenario should comprise domain-specific data collection or filtration.

6. Evaluation

In the previous section we traced the improvements of particular SMT system components. In this section we evaluate the results of the complete system, end-to-end. Currently, there are no special optimizations of our system at the ASR-MT interface. ASR output is directly piped into the MT system.

6.1. Automatic Evaluation

Automatic evaluation is given by means of BLEU [19] as in the previous sections. For sake of completeness we also provide Meteor [20] scores computed with the default settings for the corresponding target language. Table 4 summarizes the results.

In the case of pure MT our results are the same as reported last in the previous section. ASR outputs have been produced only for the Daily Dialogue domain, as this test set is also used in the next section for manual evaluation. The quality of ASR+MT drops in comparison to pure MT output which is not surprising given word accuracy and sentence-based accuracy of all systems.

6.2. Mean Opinion Score

Compared to the WMT evaluation scheme [21] that only serves as a ranking method without providing absolute quality measurements, we are interested in obtaining values that can stand on their own. Therefore for the manual evaluation part, we adapted the Mean Opinion Score (MOS) [22] method to our needs. Linguists are required to give each translation a rating from 1 (Bad) to 5 (Excellent). The MOS is the arithmetic mean of all the individual scores.

Due to the rather high similarity between the two domains we restricted the manual evaluation to the Daily Dialogue domain only. In addition to the mentioned systems the reference translation was also rated as a separate system. Annotators are presented with the original sentence and a translation result. In the case of ASR+MT the joint quality (end-to-end) is assessed.

Table 4: Automatic Evaluation and Comparison to Online-translation Systems

(a) German-to-English

DE→EN	Daily Dialogue		Travel Dialogue	
System	Bleu	Meteor	Bleu	Meteor
MT-Only	43.2	39.0	43.7	39.7
ASR+MT	22.7	24.3	–	–

(b) English-to-German

EN→DE	Daily Dialogue		Travel Dialogue	
System	Bleu	Meteor	Bleu	Meteor
MT-Only	33.1	53.2	36.0	55.0
ASR+MT	10.0	22.9	–	–

Table 5: Manual Evaluation and Comparison to Online-translation Systems

(a) German-to-English

DE→EN System	Daily Dialogue MOS	Krippendorff's α
Reference	4.54	0.301
MT-Only	4.45	0.395
ASR+MT	3.08	0.645

(b) English-to-German

EN→DE System	Daily Dialogue MOS	Krippendorff's α
Reference	4.51	0.256
MT-Only	4.38	0.317
ASR+MT	2.30	0.641

During the evaluation campaign, 20 linguists produced 12200 ratings in total. Sentences where sampled randomly without repetition from the described test sets, distributed evenly among languages, systems, and domains.

Table 5 contains the accumulated MOS results and Krippendorff's α with the interval metric [23] to assess inter-annotator agreement for all systems. Mean inter-annotator agreement for MT output across directions is 0.356. Cohen's κ calculated during the WMT evaluation campaign for MT output is usually between 0.35 and 0.40. However, the two measures may not be fully comparable due to different evaluation objectives. Mean inter-annotator agreement for ASR+MT is 0.643. By contrast, agreement seems to be lowest for reference translations, averaging at 0.279.

We suppose the difference in agreement between MT and ASR+MT output might be caused by higher variations in the latter case which makes it easier to assess quality. A failed ASR recognition will most likely result in an incomprehensible translation that will be equally penalized across annotators. We observe a good correlation between automatic and manual evaluation results.

7. Conclusions and Future Work

We demonstrated that building a high-quality MT system from publicly available parallel data is feasible and quite achievable in an unrestricted-data setting. For the chosen domain the MT-only systems reach MOS results close to the reference. Joint ASR-MT suffers currently from the ASR quality. Adding state-of-the art modifications (OSM, WCLM) to the MT systems results in significant MT quality improvements. New freely available resources for monolingual data (CC LM) make it even easier to reach good results. Our manual evaluation shows, that even BLEU results of "only" 43% for the MT component are well received by human raters.

In this paper we concentrated on the MT component, neglecting the speech-related aspects. In the current scenario, all systems communicated via 1-best results. In the future, we plan to explore n-best lists or lattice-based interfaces. The ASR components can still be improved and will most likely result in significant Speech-to-Text MT quality improvement.

8. References

[1] J. R. Smith, H. Saint-Amand, M. Plamada, P. Koehn, C. Callison-Burch, and A. Lopez, "Dirt Cheap Web-Scale Parallel Text from the Common Crawl." in *ACL*. The Association for Computer Linguistics, 2013, pp. 1374–1383.

[2] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2214–2218.

[3] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Conference Proceedings: the tenth Machine Translation Summit*, AAMT. Phuket, Thailand: AAMT, 2005, pp. 79–86.

[4] R. Steinberger, A. Eisele, S. Klocek, S. Pilos, and P. Schlter, "DGT-TM: A freely available Translation Memory in 22 languages," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), 2012.

[5] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D. Tufi, "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages," in *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006*, 2006, pp. 2142–2147.

[6] K. Wäschle and S. Riezler, "Structural and Topical Dimensions in Multi-Task Patent Translation," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), Avignon, France*, 2012.

[7] C. Buck, K. Heafield, and B. van Ooyen, "N-gram Counts and Language Models from the Common Crawl," in *Proceedings of the Language Resources and Evaluation Conference*, Reykjavík, Iceland, May 2014.

[8] F. Morbini, K. Audhkhasi, K. Sagae, R. Artstein, D. Can, P. G. Georgiou, S. Narayanan, A. Leuski, and D. Traum, "Which ASR should I choose for my dialogue system?" in *SIGDIAL*, Metz, France, 8 2013.

[9] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and*

Demonstration Sessions, ser. ACL '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 177–180.

[10] H. Johnson, J. D. Martin, G. F. Foster, and R. Kuhn, "Improving Translation Quality by Discarding Most of the Phrasetable," in *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic*, 2007, pp. 967–975.

[11] M. Junczys-Dowmunt, "Phrasal Rank-Encoding: Exploiting Phrase Redundancy and Translational Relations for Phrase Table Compression." *Prague Bull. Math. Linguistics*, vol. 98, pp. 63–74, 2012.

[12] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," in *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ser. ACL '96. Stroudsburg, PA, USA: Association for Computational Linguistics, 1996, pp. 310–318.

[13] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable Modified Kneser-Ney Language Model Estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013, pp. 690–696.

[14] N. Durrani, A. Fraser, H. Schmid, H. Hoang, and P. Koehn, "Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT?" in *ACL*. The Association for Computer Linguistics, 2013, pp. 399–405.

[15] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based N-gram Models of Natural Language," *Comput. Linguist.*, vol. 18, no. 4, pp. 467–479, 1992.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *CoRR*, vol. abs/1301.3781, 2013.

[17] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ser. ACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 160–167.

[18] M. Hopkins and J. May, "Tuning As Ranking," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1352–1362.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[20] A. Lavie and M. J. Denkowski, "The Meteor Metric for Automatic Evaluation of Machine Translation," *Machine Translation*, vol. 23, no. 2-3, pp. 105–115, Sep. 2009.

[21] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna, *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2014, ch. Findings of the 2014 Workshop on Statistical Machine Translation, pp. 12–58.

[22] International Telecommunication Union, "ITU-T Recommendation P.800.1: Mean Opinion Score (MOS) terminology," Tech. Rep., Jul. 2006. [Online]. Available: http://www.itu.int/rec/T-REC-P.800.1-200607-I/en

[23] R. Artstein and M. Poesio, "Inter-coder Agreement for Computational Linguistics," *Comput. Linguist.*, vol. 34, no. 4, pp. 555–596, Dec. 2008.