Statistical Machine Translation for Automobile Marketing Texts

Samuel Läubli[1] **Mark Fishel**[1] **Manuela Weibel**[2] **Martin Volk**[1]

[1]Institute of Computational Linguistics
University of Zurich
Binzmühlestrasse 14
CH-8050 Zürich

{laeubli,fishel,volk}
@cl.uzh.ch

[2]SemioticTransfer AG
Bruggerstrasse 37
CH-5400 Baden

manuela.weibel
@semiotictransfer.ch

Abstract

We describe a project on introducing an in-house statistical machine translation system for marketing texts from the automobile industry with the final aim of replacing manual translation with post-editing, based on the translation system. The focus of the paper is the suitability of such texts for SMT; we present experiments in domain adaptation and decompounding that improve the baseline translation systems, the results of which are evaluated using automatic metrics as well as manual evaluation.

1 Introduction

As machine translation and post-editing are gaining popularity on an industrial level, global companies turn to using their accumulated translations for setting up in-house SMT services. Substantial cost and time savings have been reported in various sectors, such as software localization (Flournoy, 2011; Zhechev, 2012) or film and television subtitling (Volk et al., 2010).

In a joint project between the University of Zurich and SemioticTransfer AG, we target a new domain: automobile marketing texts. We show that even limited amounts of in-domain translation material allow for building domain-specific SMT systems (see Section 3), and that translation quality can be significantly improved by using out-of-domain material and language-specific preprocessing (see Section 4).

Our aim is to give an example of how even small language service providers with distinct areas of specialization can successfully incorporate machine translation and post-editing into their translation workflow. In case of our project, this claim is tested on two language pairs and translation directions: German (DE) to French (FR) and to Italian (IT). A closer description of our domain is presented in the next section.

2 Domain Description

SemioticTransfer has specialized in translating marketing materials for the automobile industry. This includes print products such as brochures or price lists as well as electronic materials such as websites or newsletters, but no technical documentation such as manuals. As a whole, the domain covers a very broad spectrum of language, ranging from highly emotional and metaphorical to very dry and technical. As an example, consider the

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 265–272.

following two segments, both stemming from the same high gloss catalogue:

(a) *The R8 e-tron combines the genes of a sports coupé and a race car in an athlete's body which expresses technological progress and superiority.*

(b) *size 8.5 J x 19 at front, size 11 J x 19 at rear, with 235/35 R 19 tires at front and 295/30 R 19 tires at rear*

Tailoring SMT systems to producing good pre-translations for our entire domain is thus challenging. Before describing our corresponding experiments in Section 4, we outline the technical setup and report on the performance of simple baseline systems in our use case.

3 Baseline Translation Systems

3.1 Technical Setup

The technical setup of all experiments closely resembles the setup of the baseline systems in the shared task of the Workshop on Statistical Machine Translation (Callison-Burch et al., 2012). The only difference is that we used `IRSTLM` for language modeling (Federico et al., 2008) because of licensing issues.

We used a test set of 500 in-domain segments for automatic evaluation; these were randomly drawn from contracts that were processed after compiling the training and development sets (see section 3.2). Using a moderate test set size enabled detailed manual inspection and categorization of the machine translations, e.g., identifying the number of compounds in out-of-vocabulary (OOV) types. All automatic metric scores were calculated using `multeval` (Clark et al., 2011). Unless otherwise stated, the scores are averages over five MERT runs (Och, 2003; Bertoldi et al., 2009). METEOR scores are only given for DE–FR systems since Italian is not fully supported (Denkowski and Lavie, 2011).

As SemioticTransfer's translation workflow is based on the Across workbench[1], we implemented an RPC layer that allows for integrating Moses server instances into Across. In this way, we were able to apply various pre- and post-processing methods to translations while ensuring seamless integration of the Moses systems as a pre-translation service into SemioticTransfer's existing infrastructure.

3.2 Baseline Systems

SemioticTransfer has been using Across for several years, through which they have accumulated a lot of quality-checked translations in their translation memories. We extracted all of this material to train in-domain baseline SMT systems for both language pairs; 2,000 in-domain segments were used as a development set for tuning. Note that unlike "regular" parallel corpora, our in-domain corpus contains each translation only once, i.e., no sentence-level frequencies are available.

Using in-domain translation memory data turned out to be a promising starting point for producing good-quality pre-translations (see Table 1). Despite the small amount of training data, the systems score 33.5 (DE–FR) and 32.3 (DE–IT) BLEU points. However, the number of OOV words is high. Besides inconsistencies in punctuation and number formatting, untranslated words were found to be particularly disturbing in manual inspection of the results. As a consequence, we applied several techniques aimed at reducing the number of unknown words in our baseline systems.

4 Improved Translation Systems

As mentioned in the previous section, we focused our efforts on improving translation

[1] `http://www.across.net`

	DE–FR			DE–IT		
	In-Domain	**Europarl**	**OpenSubtitles**	**In-Domain**	**Europarl**	**OpenSubtitles**
Tokens DE	2,011,872	48,405,406	16,858,070	1,413,452	48,419,389	15,642,379
Tokens FR/IT	2,632,256	56,372,702	16,370,845	1,731,219	50,689,987	15,458,666
Tokens OOV	4.3 %	12.5 %	14.1 %	4.7 %	9.6 %	11.1 %
Segments	166,957	1,903,628	2,852,474	112,166	1,805,792	2,131,004
Tokens/Sg. DE	12.05	25.43	5.91	12.60	26.81	7.34
Tokens/Sg. FR/IT	15.77	29.61	5.74	15.43	28.07	7.25
Types DE	86,459	398,051	351,408	59,369	395,303	293,020
Types FR/IT	47,705	146,365	222,774	34,479	176,488	242,736
Types OOV	9.6 %	24.2 %	26.8 %	11.8 %	19.8 %	24.4 %
Avg. BLEU	33.5	13.5	7.5	32.3	17.3	10.4
Avg. METEOR	51.9	32.0	21.6	-	-	-
Avg. TER	51.7	68.3	79.0	55.9	65.7	73.0

Table 1: Training data for in- and out-of-domain language and translation models. OOV rates and automatic evaluation metrics refer to a test set of 500 in-domain segments (see Section 3.1).

quality by reducing the rate of OOV input types. This was done by adding general-domain corpora and combining them with our in-domain data via domain adaptation, as well as by decompounding methods on both the in-domain data and the mixed-domain set.

4.1 Related Work

Domain adaptation has been applied to most components of statistical machine translation: language models (Clarkson and Robinson, 1997; Koehn and Schroeder, 2007), word alignment (Hua et al., 2005), and translation models (Foster and Kuhn, 2007; Sennrich, 2012). Combinations of these methods often show that there is an overlap in the translation problems that the methods fix, which leads to one method "stealing" part of the effect of the others (Koehn and Schroeder, 2007; Sennrich, 2012). We perform domain adaptation with language and translation models, following the mixture-modeling approach by Foster and Kuhn (2007) and Sennrich (2012).

A common method to handle complex compounding in languages such as German is to split unknown compounds into their parts, if they occur in the training data, and join the translated parts on the output side of the translation system. The splitting can be motivated morphologically (Stymne, 2009; Hardmeier et al., 2010) or empirically (Koehn and Knight, 2003; Dyer, 2009). Also, instead of making a final decision on whether to split a compound or not, both alternatives can be passed to the translation system by representing the input text with a lattice of phrases instead of just a single sentence (Dyer et al., 2008; Dyer, 2009; Wuebker and Ney, 2012). Our approach lies between the splitting and lattice-based approaches and is especially tailored for translation between languages with heavy compounding and word order differences.

4.2 Domain Adaptation

The moderate size of our in-domain data set and relatively high OOV rates suggest adding bigger general-domain corpora to the training material. The specific nature of the in-domain data makes it necessary to use domain adaptation techniques, to avoid having the bigger general-domain data override the original

		DE–FR				DE–IT			
Metric	**Mode**	**Avg**	$\overline{s}_{\text{sel}}$	s_{Test}	p**-value**	**Avg**	$\overline{s}_{\text{sel}}$	s_{Test}	p**-value**
BLEU ↑	In-Domain Only	33.5	1.4	0.1	-	32.3	1.4	0.3	-
	Domain Adaptation (DA)	34.3	1.4	0.2	**0.00 +**	33.0	1.5	0.1	**0.00 +**
	DA + Decompounding A	35.0	1.3	0.2	**0.00 +**	33.2	1.5	0.1	0.40
	DA + Decompounding B	34.6	1.3	0.3	**0.00 –**	33.2	1.5	0.2	0.74
TER ↓	In-Domain Only	51.7	1.2	0.3	-	55.9	1.4	0.6	-
	Domain Adaptation (DA)	50.6	1.2	0.6	**0.00 +**	53.3	1.3	0.5	**0.00 +**
	DA + Decompounding A	49.5	1.2	0.4	**0.00 +**	53.1	1.3	0.3	0.53
	DA + Decompounding B	50.2	1.2	0.5	**0.00 –**	53.7	1.4	0.4	**0.00 –**
METEOR ↑	In-Domain Only	51.9	1.1	0.1	-				
	Domain Adaptation (DA)	53.0	1.1	0.2	**0.00 +**				
	DA + Decompounding A	54.3	1.1	0.2	**0.00 +**				
	DA + Decompounding B	54.3	1.1	0.3	0.80				

Table 2: Automatic evaluation of the baseline and combined systems. p-values are relative to the preceding system and indicate whether a score improves (+) or decreases (–) significantly.

domain-specific translations.

We used two freely available out-of-domain corpora for these experiments: Europarl v7 (Koehn, 2005) and OpenSubtitles (Tiedemann, 2009). Stand-alone systems trained on these corpora result in unusable translations with very low scores (see Table 1).

For domain adaptation, we used multi-domain mixture-modeling (Foster and Kuhn, 2007) for language and translation models. The main distinctive feature of that approach is that instead of a binary in-domain/out-of-domain treatment of the data sets, each separate domain is assigned a weight, reflecting its similarity to in-domain (parallel) texts. Mixture-modeling and optimization of these weights is implemented in `IRSTLM` for language models (Federico et al., 2008) and in `tmcombine` for translation models (Sennrich, 2012).

The weight distribution discovered by the optimization step for both kinds of models and both language pairs was very similar: around 93% distribution mass to the in-domain data, 5-6% to Europarl and 1-2% to OpenSubtitles. Resulting translation quality estimation scores are presented in Table 2; all scores for both language pairs show a stable and significant improvement. The OOV rate for types dropped from 9.6% to 5.8% for DE–FR and from 11.8% to 6.1% for DE–IT in the adapted model combination. Manual evaluation and a more detailed description of the domain-adaptation experiments can be found in (Läubli et al., 2013).

4.3 Decompounding

Categorizing the remaining German OOV types in our DE–FR test set revealed that 65% of them were either nominal or adjectival compounds. While adding out-of-domain data reduced the number of "normal" OOV words[2] by 81%, only 24% of the OOV compounds could be translated in this way. In other words, our categorization confirmed that even big parallel corpora are sparse of domain-specific German compounds such as *Fahrzeugmodells* (*car type*, genitive), but also general-domain compounds such as *Winterbeginn* (*onset of winter*).

[2] We defined normal words as the set of types that do not belong to any of the following categories: compounds, numbers, foreign words, spelling errors, casing errors, tokenization errors.

4.3.1 Basic Approach

Koehn and Knight (2003) have addressed compounding in machine translation with an empirical method. Their approach builds upon the fact that even if a compound is out-of-vocabulary, the training data or phrase table might still contain its distinct parts. Koehn and Knight thus consider all splits S of an unknown compound C into known parts p (substrings) that entirely cover C. Given the count of words in the training corpus, the best split $\hat{s}$ has the highest geometric mean of word frequencies of its n parts:

$$\hat{s} = \arg\max_{S} \Big(\prod_{p_i \in S} \text{count}(p_i) \Big)^{\frac{1}{n}}$$

This leaves compounds unbroken if they appear more frequently in the training material than their parts. For example, *Fahrzeugmodell* (*car type*) is preserved in our DE–FR model, while *Fahrzeugmodells* (*car type*, genitive) is split into *Fahrzeug Modells*.

We applied Koehn and Knight's decompounding method to our domain-adapted DE–FR and DE–IT systems. Using the corresponding Moses implementation, we split compounds on the German side of each corpus (in-domain, Europarl, and OpenSubtitles) to train new translation models. This leads to a significant improvement of all metric scores for DE–FR, but not for DE–IT (see Table 2, DA + Decompounding A).

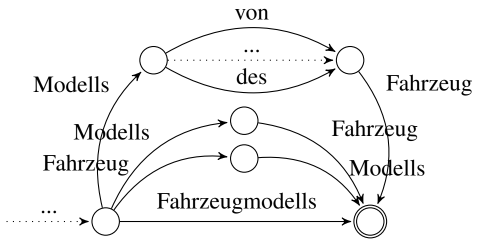
The actual translations of compounds in our test set did not meet our expectations. Despite the metric improvements, we identified two severe problems: missing function words and word order. In both French and Italian, the parts of a compound are often connected by function words such as *de* (FR) or *di* (IT). Moreover, their order is usually reversed, at least in compounds with two stems only. Consider for example the German compound *Fahrzeugmodells* (*car type*, genitive), which translates into *modèle de véhicule* (FR) and *modello di veicolo* (IT), respectively.

As function words and reordering are crucial for translating German compounds into romance languages such as French or Italian, we propose a modified approach to decompounding in the following section.

4.3.2 Using Ambiguous Input

Dyer (2009) and Hardmeier et al. (2010) have used word segmentation lattices to pass multiple analyses of an input segment to an SMT decoder. The intuition behind this concept is to avoid the propagation of errors caused by selecting one single best hypothesis and to let the multiple splits be evaluated against each other at runtime.

The lattice-based approach, however, does not address the changed order or insertion of function words, which is what we propose. Our updated input lattice includes the original compound, a path of its two split parts as well as a similar path but with the order of compound parts switched, and finally the switched-order path with several function word alternatives inserted in between. The latter are inserted in the source language (German) to avoid forcing certain translation variants. For *Fahrzeugmodells*, the modified input lattice would look like this:



At this experimental stage, we have weighted all paths equally. Also, the choice of function words is naturally arbitrary.

To test the modified approach, we first ran a pilot experiment on in-domain data only for DE–FR (see Table 3). Although the automatic

Mode	BLEU	METEOR	TER
In-Domain Only	33.5	51.9	51.7
+ Decompounding A	34.3	53.2	50.0
+ Decompounding B	34.4	53.3	50.5

Table 3: Effects of decompounding techniques (see Sections 4.3.1 and 4.3.2) on DE–FR systems using in-domain training material only.

scores of the Koehn and Knight (2003) approach (Decompounding A) and our lattice-based extension (Decompounding B) did not differ significantly in terms of automatic metrics, we found a number of compounds in our test set that were now translated correctly (*modèle de véhicule* instead of *véhicule modèle*) or in a more comprehensible way (*la liste des amis* instead of *les amis liste*).

By adding more data through domain adaptation (see Section 4.2) however, Decompounding B is outperformed by Decompounding A (see Table 2). In order to assess whether this difference is also conceivable for an actual translator, we conducted a human evaluation experiment.

4.3.3 Human Evaluation

Our human evaluation was primarily aimed at testing if translators or post-editors prefer machine translations that involve decompounding over such that leave unknown compounds untranslated. The human evaluator, fluent in the target languages (FR, IT) as well as familiar with the specific domain terminology, compared the decompounding setups (Decompounding A and B) to No Decompounding in both language pairs. This resulted in four tasks, each of which consisted in comparing 150 segments. This procedure—commonly referred to as pairwise ranking—is well-established in MT research (Callison-Burch et al., 2012). We included between four and twenty duplicates per task in order to measure the intra-annotator agreement, which turned out to be $\kappa = 0.93$, i.e., “almost perfect” according to Landis and Koch (1977). We also calculated a p-value for each task in order to quantify genuine differences between each two systems. As in (Callison-Burch et al., 2012), we ignored ties and applied the Sign Test for paired observations.

In contrast to the automatic evaluation results, the human evaluation shows a better performance of Decompounding B, which outperformed No Decompounding in both language pairs, whereas Decompounding A only performed better for German–Italian. However, the differences are not statistically significant.

Due to the domain-specific training material, frequently used automobile and marketing terms such as *LED-Leseleuchten* (*LED reading lights*) → *lampes de lecture à DEL* (DE–FR) or *Bremsenergie-Rückgewinnung* (*recovery of breaking energy*) → *il recupero dell’ energia di frenata* (DE–IT) were translated correctly even by the systems with no decompounding; only unknown compounds could not be translated. Thus, the human evaluator’s preference for translations that involve decompounding is obvious only in cases of successful translations of unknown compounds, such as *Datenschutzerklärung* (*privacy statement*) → *déclaration de la protection des données* (DE–FR; Decompounding B).

5 Conclusion

In this paper, we have shown how translation memories of limited size can be used to build domain-specific SMT systems for automobile marketing texts. Our work is targeted at enabling language service providers with expertise in distinct domains and language pairs to incorporate post-editing into their translation workflow.

	DE–FR					DE–IT				
Mode	**Sg.**	**loss**	**tie**	**win**	**p-val.**	**Sg.**	**loss**	**tie**	**win**	**p-val.**
Decompounding A	146	.18	.68	.14	.46	130	.20	.56	.24	.60
Decompounding B	138	.18	.54	.28	.08	140	.24	.46	.30	.42

Table 4: Human Evaluation. Both decompounding methods (see Sections 4.3.1 and 4.3.2) are evaluated against the No Decompounding baseline (see Section 4.2). p-values indicate significant differences between two systems ($win > loss$). Intra-Annotator Agreement $\kappa = 0.93$.

In particular, we have used freely available out-of-domain data to reduce the number of untranslated words (OOV) in our translation models. Combining in- and out-of-domain resources into weighted mixture-models (Sennrich, 2012) ensures that domain-specific terminology prevails over alternative translations. The OOV rate was further reduced by splitting compounds in German source segments; in our German–French test set, the number of OOV types dropped from 194 to 60 (-69%) through domain adaptation and decompounding. Altogether, our combined systems outperform the in-domain only baselines significantly in both language pairs in terms of BLEU, METEOR, and TER.

Acknowledgements

This work was supported by Grant No. 11926.2 PFES-ES from the Swiss Federal Commission for Technology and Innovation.

References

Nicola Bertoldi, Barry Haddow, and Jean-Baptiste Fouet. Improved minimum error rate training in moses. *The Prague Bulletin of Mathematical Linguistics*, 91(1):7–16, 2009.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of WMT*, pages 10–51, Montréal, Canada, 2012.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of ACL/HLT*, pages 176–181, Portland, USA, 2011.

Philip R. Clarkson and Anthony J. Robinson. Language model adaptation using mixtures and an exponentially decaying cache. In *Proceedings of ICASSP*, volume 2, pages 799–802, Munich, Germany, 1997.

Michael Denkowski and Alon Lavie. Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of WMT*, pages 85–91, Edinburgh, UK, 2011.

Chris Dyer. Using a maximum entropy model to build segmentation lattices for mt. In *Proceedings of HLT-NAACL*, pages 406–414, Boulder, USA, 2009.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. Generalizing word lattice translation. In *Proceedings of ACL*, pages 1012–1020, Columbus, USA, 2008.

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

Raymond Flournoy. MT use within the enterprise: Encouraging adoption via a unified MT API. In *Proceedings of MT Summit XIII*, pages 234–238, Xiamen, China, 2011.

George Foster and Roland Kuhn. Mixture-model adaptation for SMT. In *Proceedings of WMT*, pages 128–135, Prague, Czech Republic, 2007.

Christian Hardmeier, Arianna Bisazza, and Marcello Federico. FBK at WMT 2010: word lattices for morphological reduction and chunk-based reordering. In *Proceedings of WMT/MetricsMATR*, pages 88–92, Uppsala, Sweden, 2010.

Wu Hua, Wang Haifeng, and Liu Zhanyi. Alignment model adaptation for domain-specific word alignment. In *Proceedings of ACL*, pages 467–474, Ann Arbor, USA, 2005.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand, 2005.

Philipp Koehn and Kevin Knight. Empirical methods for compound splitting. In *Proceedings of EACL*, pages 187–193, Budapest, Hungary, 2003.

Philipp Koehn and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of WMT*, pages 224–227, Prague, Czech Republic, 2007.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

Samuel Läubli, Mark Fishel, Martin Volk, and Manuela Weibel. Combining domain-specific translation memories with general-domain parallel corpora in statistical machine translation systems. In *Proceedings of NoDaLiDa*, pages 331–341, Oslo, Norway, 2013.

Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167, Sapporo, Japan, 2003.

Rico Sennrich. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of EACL*, pages 539–549, Trento, Italy, 2012.

Sara Stymne. Compound processing for phrase-based statistical machine translation. Master's thesis, Linköping University, Linköping, Sweden, 2009.

Jörg Tiedemann. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Proceedings of RANLP*, volume V, pages 237–248, Borovets, Bulgaria, 2009.

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström. Machine translation of TV subtitles for large scale production. In *Second Joint EM+/CNGL Workshop*, pages 53–62, Denver, USA, 2010.

Joern Wuebker and Hermann Ney. Phrase model training for statistical machine translation with word lattices of preprocessing alternatives. In *Proceedings of WMT*, pages 450–459, Montreal, Canada, 2012.

Ventsislav Zhechev. Machine Translation Infrastructure and Post-editing Performance at Autodesk. In *Proceedings of WPTP*, pages 87–96, San Diego, USA, 2012.