Modeling verbal inflection for English to German SMT

Anita Ramm
IMS
University of Stuttgart, Germany
ramm@ims-uni.stuttgart.de

Alexander Fraser
CIS
University of Munich, Germany
fraser@cis.uni-muenchen.de

Abstract

German verbal inflection is frequently wrong in standard statistical machine translation approaches. German verbs agree with subjects in person and number, and they bear information about mood and tense. For subject–verb agreement, we parse German MT output to identify subject–verb pairs and ensure that the verb agrees with the subject. We show that this approach improves subject-verb agreement. We model tense/mood translation from English to German by means of a statistical classification model. Although our model shows good results on well-formed data, it does not systematically improve tense and mood in MT output. Reasons include the need for discourse knowledge, dependency on the domain, and stylistic variety in how tense/mood is translated. We present a thorough analysis of these problems.

1 Introduction

Statistical machine translation of English into German faces two main problems involving verbs: (i) correct placement of the verbs, and (ii) generation of the appropriate inflection for the verb.

The position of verbs in German and English differs greatly and often large-range reorderings are needed to place the German verbs in the correct positions. Gojun and Fraser (2012) showed that the *preordering* approach applied on English–to–German SMT overcomes large problems with both missing and misplaced verbs.

Fraser et al. (2012) proposed an approach for handling inflectional problems in English to German SMT, focusing on the problems of sparsity caused by nominal inflection. However, they do not handle the verbs, ensuring neither that verbs appear in the correct position (which is a problem due to the highly divergent word order of English and German), nor that verbs are correctly inflected (problematic due to the richer system of verbal inflection in German). In many cases, verbs do not match their subjects (in person and number) which makes understanding of translations difficult. In addition to person and number, the German verbal inflection also includes information about tense and mood. If these are wrong (i.e. do not correspond to the tense/mood in the source), very important information, such as point of time and modality of an action/state expressed by the verb, is incorrect. This can lead to false understanding of the overall sentence.

In this paper, we reimplement the nominal inflection modeling for translation to German presented by Fraser et al. (2012) and combine it with the reordering of the source data (Gojun and Fraser, 2012). In a novel extension, we present a method for correction of the agreement errors, and an approach for modeling the translation of tense and mood from English into German. While the subject-verb agreement problems are dealt with successfully, modeling of tense/mood translation is problematic due to many reasons which we will analyze in detail.

In Section 2, we give an overview of the processing pipeline for handling verbal inflection. The method for handling subject–verb agreement errors is described in Section 3, while modeling of tense/mood translation is presented in Section 4. The impact of the proposed methods for modeling verbal inflection on the quality of the MT output is shown in Section 5. An extensive discussion of the problems related to modeling tense/mood is given in Section 6. Finally, future work is presented in Section 7.

Proceedings of the First Conference on Machine Translation, Volume 1: Research Papers, pages 21–31,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

2 Overall architecture

2.1 Ensuring correct German verb placement

Different positions of verbs in English and German often require word movements over a large distance. This leads to two problems in German translations generated by SMT systems concerning the verbs: either the verbs are not generated at all, or they are placed incorrectly.

To ensure that our MT output contains the maximum number of (correctly placed) finite verbs, we reorder English prior to training and translation using a small set of reordering rules originally described by Gojun and Fraser (2012). The verbs in the English part of the training, tuning and testing data are moved to the positions typical for German which increases the syntactic similarity of English and German sentences. We train an SMT system on the reordered English and apply it to the reordered English test set.

This approach has good results in terms of the position of the verbs in German translations. However, the problem of incorrect verbal inflection is unresolved. In fact, the reordering makes the agreement problems even worse due to movements of verbs away from their subjects (cf. Section 3.1).

2.2 Inflection of the German SMT output

Fraser et al. (2012) proposed a method for handling nominal inflection for English to German SMT. They work with a *stemmed* representation of the German words in which certain morphological features such as case, number, etc. are omitted. After the translation step, for nominal stemmed words in the MT output, morphological features are predicted using a set of pre-trained classifiers and finally surface forms are generated resulting in fully-inflected German MT output.

In their approach, the verbs are neither stemmed nor inflected, but instead handled as normal words. Thus, in the translation step, the decoder (in interaction with the German language model) decides on the inflected verb forms in the final MT output.

2.3 Adding verbal inflection modeling

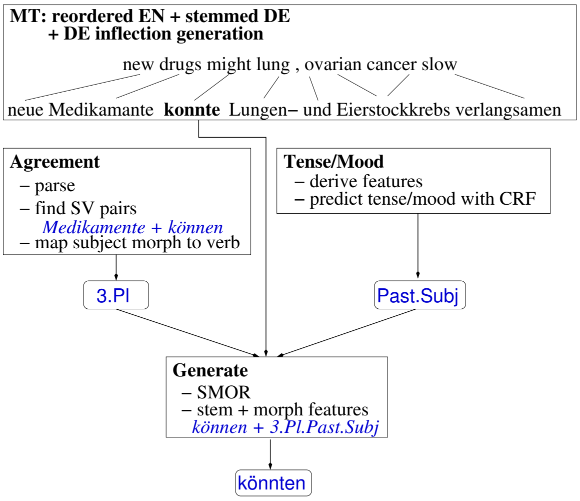
MT: reordered EN + stemmed DE
+ DE inflection generation
new drugs might lung , ovarian cancer slow
neue Medikamante konnte Lungen– und Eierstockkrebs verlangsamen
Agreement
– parse
– find SV pairs
Medikamente + können
– map subject morph to verb
Tense/Mood
– derive features
– predict tense/mood with CRF
3.Pl
Past.Subj
Generate
– SMOR
– stem + morph features
können + 3.Pl.Past.Subj
könnten

Figure 1: Processing pipeline. The verbal inflection modeling consists of two components: (i) a component for deriving agreement features person and number, and (ii) a component for predicting tense and mood. The inflected verbs are generated with SMOR (Schmid et al., 2004), a morphology generation tool for German.

As a baseline SMT system, we use a system trained on the reordered English sentences (cf. Section 2.1) and stemmed German data with nominal inflection modeling as a post-processing step (cf. Section 2.2). In our system, we extend the baseline by identifying finite verbs in the baseline MT output, predicting their morphological features and finally producing the correct inflected output (see Figure 1).

Verbal morphological features include information about person/number, as well as tense and mood. Particularly the modeling of tense/mood translation is interesting: in this paper, we present a method to model the translation of English tense and mood into German considering all German tenses/moods in a single model. In addition, we present a detailed discussion which is, to our knowledge, the first deep analysis of this topic.

The processing pipeline is given in Figure 1. After translation of the reordered English input to a German stem-like representation, the nominal feature prediction is performed followed by our novel verbal feature prediction. Finally, the entire German MT output is inflected by combining the stems and the predicted features to produce surface forms (normal words).

3 Correction of the subject–verb agreement

3.1 Problem description

In many languages, the subject is located near the corresponding finite verb. However, in languages such as German, the subject might be very far from

Data	avg dist in words	>5 words
News	3.9	24%
Europarl	3.7	22%
Crawled	2.9	15%

Table 1: Subject–verb distances in German texts.

the verb. We extracted subject–verb pairs from German corpora and computed their distances. The results are summarized in Table 1.

News and *Europarl* are composed of more complex sentences than the corpus crawled from the internet. While in the crawled data, there are more sentences with smaller subject–verb distances, *News* and *Europarl* expose larger distances between subjects and finite verbs.

Although the average distance in words is rather small, there is a fair amount of subject–verb pairs with distance larger than 5 words (in Europarl 22%, in News 25%) which are problematic for training the translation system. Even for small distances, it is not guaranteed that the agreement is generated correctly due to the missing appropriate translation phrases. Moreover, the German language model trained on the same data would probably have problems to extract n-grams which ensure the correct subject–verb agreement for *all* possible subject–verb combinations.

Translating reordered English (cf. section 2.1) dramatically improves the problems of misplaced and missing verbs, but at the same time makes the extraction of translation phrases with subject–verb agreement even harder. Particularly problematic are movements of the verbs in subordinate clauses where the entire German VP is placed at the clause end, while the subject is normally placed in the 2nd position (after the complementizer). In our training data, 20% of the clauses are reordered in a way that the distance between the reordered finite verb and the subject is more than 5 words.

An example of a reordered English subordinate clause is given in Figure 2: the English verb *said* is ambiguous with respect to person and number. Translated independently from its subject, it is not guaranteed that the German translation will contain the correctly inflected finite verb since the German language model is very unlikely to have the exact 6-gram which could ensure the agreement between the subject *ich/I* and the inflected verb *habe/have*.

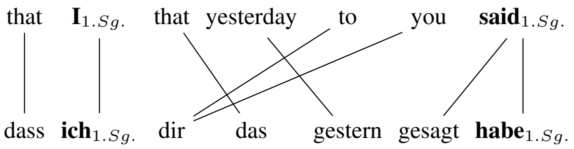

Figure 2: Example of a subject–verb distance caused by the reordering of the English clause *'that I said that yesterday to you'*.

3.2 Parsing for detection of subject–verb pairs

Agreement correction depends on correct identification of subject–verb pairs. Although we work with English parses where the subjects can be correctly identified in many cases, this information source seems not to be sufficient. Problematic are syntactic divergences where the English subject does not correspond to the German subject.

Initially, we aimed at predicting agreement features. However, we were not able to build a classifier with satisfying results due to the problems mentioned above. We thus applied a method implemented in *Depfix* (Rosa et al., 2012). They parse the MT output, extract subject–verb pairs from the trees and copy the agreement information of the subject to the corresponding verb. Although the idea of parsing MT output may not sound very promising, the results are surprisingly good.

We implement the agreement correction for English–German SMT as an automatic post-editing step applied on the fully inflected MT output. The MT output is first annotated with morphological information (Müller et al., 2013) and subsequently parsed (Björkelund and Nivre, 2015). The person and number of the subjects are then mapped from the subject to the finite verbs.

To generate the appropriate inflected verb, we use SMOR (Schmid et al., 2004), a morphology generation tool for German. Based on the stem of the verb, as well as its morphological features person, number, tense and mood (cf. section 4), the inflected verb form is generated. In case the tool produces multiple surface form possibilities (which is very rare for verbs) we use the frequency of the alternatives (derived from a large German corpus) as a filter: the most frequent alternative is chosen.

4 Modeling tense and mood

We define the modeling of tense and mood as a classification problem. In the following, we present the problem in more detail, motivate the machine learning features that we use and give a detailed evaluation of the classification model.[1]

4.1 Problem description

We distinguish between tense/mood of the finite verbs and tense/mood of the clauses. The German finite verbs can be present or past. As for the mood, they can be indicative, subjunctive and imperative.[2]

4.1.1 Tense

The tense of the finite verb does not necessarily match the clausal tense. For example, given the clausal tense *perfect*, the finite auxiliary is in present tense, while the main verb is a past participle: $[habe_{Pres.Ind}$*/have* $gesagt_{ppart}/said]_{perfect}$.

We model the translation of clausal tenses from English to German and than map the clausal tense to the corresponding tense of the finite verb.

German has six indicative clausal tenses (cf. Table 3). While in some languages, the use of tense underlies strict rules, the use of tenses in German often follows from the register (spoken vs. written) or even from the author's stylistic preferences (e.g. (Sammon, 2002), (Collins and Hollo, 2010)).

4.1.2 Mood

In addition to six indicative German tenses, we also distinguish two further tense/mood combinations: *Konjunktiv I* (present subjunctive) and *Konjunktiv II* (past subjunctive). While Konjunktiv II corresponds to English conditionals, Konjunktiv I is used in the context of indirect speech.

The use of subjunctives in German is not only quite complex, but also largely user- and register-dependent. For example, while Konjunktiv I occurs in *Europarl* and *News*, it is almost never used in the web-crawled corpus, as we will see in the following sections.

Info type	Example
STEMS	$haben_{VAFIN}$ $sagen_{VVPP}$
POS	VAFIN, VVPP
RFTagger	1.Sg.Past.Subj
RULE	if VP consists of an auxiliary (VAFIN) and a participle (VVPP) and if the finite verb is Past.Subj ⇒ konjunktivII (past subjunctive)'

Table 2: Information used to derive tense for the VP *hätte/would-have gesagt/said.*

4.2 Tense/mood prediction model

4.2.1 Model

For the classifier training, we use the toolkit *Wapiti* (Lavergne et al., 2010) which supports both multi-label maximum entropy classification and bigram linear-chain CRF classification.

We train a maximum entropy model, as well as a bigram linear-chain CRF model. The latter model captures intra-sentence tense/mood dependencies, i.e. between verbs within clauses of a single sentence: the prediction of tense/mood for the current clause considers the prediction made for the preceding clause.

Inter-sentence dependencies are however not modeled. The prediction for the first clause of the sentence under consideration does not take the last prediction made for the previous sentence into account.

4.2.2 Data

The training instances are extracted from Europarl, News Commentary and Crawled corpus. The English part of the corpus is parsed with the constituent parser of (Charniak and Johnson, 2005), while the German data is stemmed (cf. Section 2.2). We use the automatically computed word alignment (Och and Ney, 2003) in order to identify verb pairs in a given sentence pair.

We work with a set of 8 labels which includes six German tenses and the two subjunctive moods (see Table 3). In the training data, the labels are annotated by rule-based mapping of the German VPs. We use information about the verbs, their POS tags, as well as the morphological analysis of the finite verb to derive labels for each German VP (see Table 2 for an example mapping). The distribution of the labels in the corpora we use is given in Table 3.

For each finite verb, a training instance with features from English and German parallel sentence is extracted. Finite verbs of a sentence build

[1] Note that aspect is not encoded in the German verbal morphology. For expressing progressive aspect, adverbials (e.g. *gerade/at the moment*) or prepositional phrases (e.g. *Ich/I bin/am am/at Arbeiten/work 'I am working')* are used (cf. e.g. (Heinold, 2015)). In this work, we do not explicitly model aspect.

[2] In this work, we ignore imperatives. Imperatives do not bear morphological information about tense and mood: they solely distinguish the person (singular/plural). We simply retain imperatives generated by the baseline system.

tense/mood	news	europarl	crawl	news+ euro+ crawl
present	54	63	71	62
perfect	11	14	12	12
imperfect	19	6	9	11
pluperfect	3	2	3	2.6
future I	1	3	1	1.6
future II	0.5	0.1	1	0.5
konjunktiv I	1	0.9	0.7	0.8
konjunktiv II	8	7	2	5.8

Table 3: Distribution of the tense/mood labels in the German corpora (given in percentage).

a sequence which allows for taking into account the tense/mood dependency between finite verbs within a sentence.

For the classifier training, we only use instances where the German verb is aligned with at least one English word. Furthermore, if the mapping of a VP to tense in one of the languages fails, the training instance is omitted as well. In total, we extract 5.2 million training instances.

4.2.3 Feature set

Each German finite verb gets features assigned from both English and German. The English features are extracted on the basis of the clauses. Given the alignment between the German finite verb and a specific word in English, the features are used which are extracted from the clause the English word is placed in. Since in the training, finite German verbs may be aligned with arbitrary English words (i.e. not only verbs), the clause-wide features allow to extract features also for these verbs.

Lexical features Lexical features give information about lexical choice of the verbs. To avoid sparsity problems, we abstract the English VP to a certain extent: we use information about (i) main (meaning-bearing) verbs, (ii) a sequence of auxiliaries without the main verb since the auxiliaries in English are used to form different tense/moods. By having access to the main verbs from both the current clause, as well as from the preceding clause, we account for the fact that the verbs (or their sequences) influence the use of tense/mood.

Contextual features Words preceding the German finite verb are useful for some specific contexts in which Konjunktiv is used.

Semantics/discourse The combination of clauses, i.e. clause types, has impact on the choice of tense/mood. Moreover, we use the information whether the sentence is composed (i.e. consists of more than one clause) to account for the fact that some tense/moods, e.g. Konjunktiv, are rarely used in simple sentences. The conditional context is derived by a simple check whether the conjunction in the subordinate clause is *if*.

Feature	English	German
finite verb	said	haben
finite verb align	–	said
VP	said	–
VP correct	yes	–
main verb	said	sagen
prev. clause main verb	–	denken
auxiliaries	VBD	–
main suffix	id	–
sentence main verb	think	–
word-1	–	gesagt
word-2	–	gestern
clause type	SBAR	–
preceding clause type	S-MAIN	–
following clause type	END	–
syntactical tense	past	–
logical tense	past	–
conditional context	no	–
composed sent	yes	–

Table 4: Full feature set for modeling tense/mood translation. The values are derived for the German finite verb *haben/have* from the clause pair given in Figure 2 assuming that the full English sentence is *'I think that I said that yesterday to you.'*

The features are summarized in Table 4. Our model does not only use these features, but also a number of their combinations to strengthen contexts for specific tense/moods.

4.2.4 Classifier evaluation

Although both maximum entropy, as well as CRF models trained on the same data using the same feature set perform equally well, CRF performs better for certain labels as shown in Table 5.

We further evaluate the CRF model on test sets from different domains (cf. Table 6). Note that the test sets are well-formed sentences taken from the corpora we work with. We contrast evaluation results gained on well-formed test data to those obtained for noisy MT output. The evaluation on the well-formed data is given in F_1-scores while the MT output is evaluated with BLEU.

The row *mostFreqTense* is considered to be a baseline: the verbs are annotated with tense which is the most frequent German tense given a specific English tense (cf. Figure 3). It is interesting that

tense/mood	F_{1CRF}	F_{1me}
present	0.92	0.92
perfect	0.81	0.81
imperfect	0.85	0.85
pluperfect	**0.74**	0.73
future I	**0.84**	0.83
future II	0.50	0.50
konjunktiv I	**0.27**	0.17
konjunktiv II	0.83	0.83
overall	0.87	0.87

Table 5: Performance of a CRF vs. maximum entropy classifier gained for a test set containing 5,000 sentence from the news corpus.

the baseline performs equally well when applied on news and crawl, it however leads to lower F_1 for the europarl test set. This indicates that the tense usage in europarl deviates from that in news and crawled corpora.

Our model is considerably better than the baseline. It leads to better results on both well-formed test sets, as well as on the MT output.

tense/mood	F_{1CRF}			BLEU
	news	europarl	crawl	MT-news
mostFreqTense	0.70	0.64	0.70	21.79
our model	0.87	0.90	0.88	21.95

Table 6: Classifier evaluation using different features and different test sets. Each of the clean data test sets contain 5,000 sentences. Clean data sets are evaluated in terms of F_1 scores, while the MT output is evaluated with BLEU.

The difference in performance gained on test sets from different domains (although small) raises the question whether the classifier is solely to be trained on in-domain data. Since we work with MT output of the news test set, we would have to train the classifier only on the news data. Due to the corpus size (272k sentences), we get into sparsity problems since many lexical features are used. A further reason for using additional (out-of-domain) training data are low-frequent labels which then get more training instances.

In summary, the evaluation indicates that a single classifier leads to different results when applied on data from different domains. Furthermore, the initial experiments showed that having better results on the clean data does not necessarily lead to better results for the noisy MT output.

5 Verbal morphology in MT output

5.1 Baseline system

Our baseline system is trained on reordered English sentences (cf. Section 2.1) and stemmed German data (cf. Section 2.2). It is trained on a corpus consisting of 4.5 M sentences from news, Europarl and crawled texts. It uses a 5-gram language model trained on 1.5 billion German words.

The baseline system translates reordered English into stemmed German in which the verbs are surface forms and enriched with POS tags.

5.2 Evaluation of the verbs in MT output

The baseline SMT system is applied on a news test set from WMT 2015.[3]

The baseline MT output we aim at correcting is surprisingly good. The stem- and surface-based comparison of the verbs in the baseline with the reference revealed that 82% of the verbs in the baseline are already correctly inflected. This quite high number though takes only 21% of the verbs in the baseline into account: nearly 80% of the verbs in the baseline do not match the reference, i.e. the lexical choice (the lemma) of the verbs differs from the reference.

Our verbal inflection correcting system changes 242 (6%) of the verbs output by the baseline SMT system. Given the strong baseline we work with, we would in fact do worse if we changed more (i.e. already correctly inflected) verbs.

Considering the fact that most of the finite verbs do not match the reference and are thus not considered with automatic metrics such es BLEU (cf. Section 5.2.1), we also carried out a human evaluation which is presented in Section 5.2.2.

5.2.1 Automatic evaluation

In Table 7, the BLEU scores (Papineni et al., 2002) of the MT output with predicted verbal inflection are presented.

	BLEU_{ci}
Surface	21.59
Baseline	22.00
Verbal inflection	22.05
Agreement	22.08
Tense/mood	21.95

Table 7: BLEU scores of MT outputs with corrected verbal inflection.

[3] http://www.statmt.org/wmt15/

Verbal inflection denotes MT output for which all verbal features are derived/predicted and then used to generate the inflected verb forms. The translation quality does not increase (in terms of BLEU) significantly. Most of the improvement comes from the agreement correction (given in row *Agreement*) while the tense/mood prediction (row *Tense/mood*) lowers the BLEU score.

5.2.2 Manual evaluation of MT

70 sentence pairs consisting of the baseline MT output and MT output with corrected verbal inflection with respect to tense and mood were evaluated by four human evaluators. The evaluators annotated the better translation alternative with 1, the worse one with 2. For each of the translations, the *majority vote* (most frequent annotation) was computed. The counts of the human votes are given in Table 8.

	Grade			
MT	1	2	3	nA
Baseline	**29**	19	4	19
Verbal inflection	17	31	4	19

Table 8: Results of human evaluation. 1 = better, 2 = worse, 3 = don't know, nA = no majority vote.

Human evaluators prefer the choice of tense (expressed in verbal inflection) made by the baseline. Only a third of the alternatives with verbal inflection handling are considered to be better than the baseline. An interesting fact is that the annotator agreement in terms of Kappa was only 0.33 which means that the annotators often disagreed which translation alternative was better.

In Table 9, a few example MT outputs are shown in which the verbal inflection is correct, while the baseline is incorrect. The VI translation of SRC1 shows corrected agreement between the plural subject *Kläger/claimants* and the finite verb *legten/presented*. The translations of SCR2 and SRC3 show the corrected tense. In SRC2, the English verb in past tense is in VI also translated as past tense. In SRC3, the German translation of the subordinate clause should be past subjunctive as generated by VI.

VI correct	SRC1	the claimants presented proof of extortion
	BL	*legte$_{3.Sg}$ die Kläger Beweise von Erpressung
	VI	legten$_{3.Pl}$ die Kläger Beweise von Erpressung
	SRC2	then he put his finger on it
	BL	dann *legt$_{Pres.Ind}$ er seinen Finger auf sie
	VI	dann legte$_{Past.Ind}$ er seinen Finger auf sie
	SRC3	I fear I may need more surgery
	BL	ich fürchte, ich *kann$_{Pres.Ind}$ eine Operation nötig
	VI	ich fürchte, ich könnte$_{Past.Subj}$ eine Operation nötig
	SRC4	Maybe his father intended to be cruel
VI incorrect	BL	vielleicht soll$_{Pres.Ind}$ seine Vater grausam zu sein
	VI	vielleicht *sollte$_{Past.Subj}$ seine Vater grausam zu sein
	SRC5	" i have rung mr piffl and suggested that we get together "
	BL	"ich habe$_{Pres.Ind}$ geklingelt Herr piffl und schlug vor, dass wir gemeinsam"
	VI	"ich *hatte$_{Past.Ind}$ geklingelt Herr piffl und schlug vor, dass wir gemeinsam"
	SRC6	no word could get beyond the soundproofing
	BL	kein Wort konnte üer die Schalldämmung
	VI	kein Wort *könte über die Schalldämmung

Table 9: Example of MT outputs with improved (upper part) and incorrect verbal inflection (lower part). *SRC* denotes the source sentences, the baseline translations are indicated with *BL*, while the translations with verbal inflection handling are indicated with *VI*.

The VI translation of *intended* in SRC4 retains the tense in the source sentences. The human evaluators, however, prefer the baseline translation, which switches to present tense. German has two past tenses: the baseline translation of *have rung* in SRC5 is perfect (*habe geklingelt*), while the VI translation is pluperfect (*hatte geklingelt*). Even for a human, it is hard to decide which of the translations is *better*. The translation of SRC6 shows a problem with English modal verbs such as *could* which expose functional ambiguity. As subjunctive, *could* almost always translates into subjunctive German modal *könnte*. Thus the model always predicts konjunktiv II given English modals for which the past indicative form equals to the subjunctive form.

6 Discussion

6.1 Subject–verb agreement

Correction of the subject–verb agreement proposed by Rosa et al. (2012) and adapted in this work for English–German SMT, relies on how accurate the identification of the subject–verb rela-

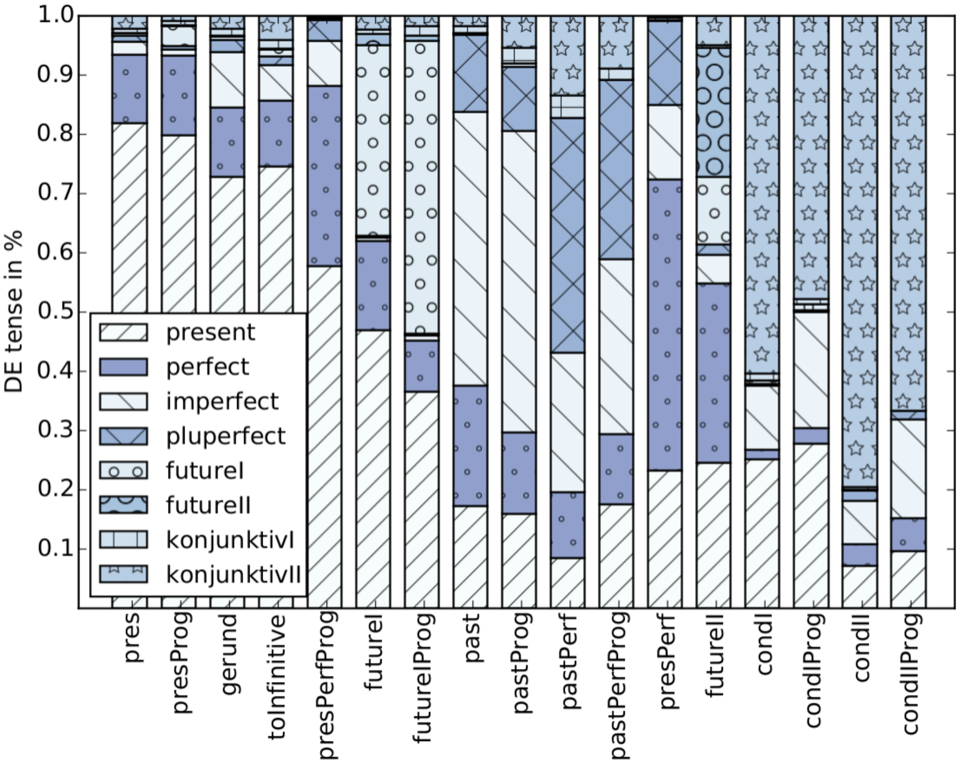

Figure 3: Distribution of tense translations derived from the training corpora (news, europarl, crawl). English tense/mood values are given on the x-axis, while the percentage of the German tense/moods for the corresponding EN tense/mood is given on the y-axis.

tions in noisy MT output is. The better the translation, the higher the probability of acquiring correct subject–verb pairs from the parse trees. However, the quality of the translations varies greatly, even within a single test set. Rosa et al. (2012) reported on different results achieved for different test sets. Another possibility is to use a classification model which predicts agreement features of the verbs using various contextual information as successfully applied on English–Spanish (Gispert and Mariño, 2008).

Our attempt to build such a model for German, led to disappointing results: on the one hand, a more accurate identification of the subjects in the English constituent parse trees is required: the use of the dependency trees combined with pronoun resolution (similar to a simple pronoun resolution described in (Avramidis and Koehn, 2008)) might reduce this problem. More correct subject identification in the source language is however not sufficient: due to syntactic divergences, the German subject may match other constituents in the source language (e.g. object or preposition phrase). A prediction model having access to information extracted from both English dependency trees, as well as German MT parses (in combination with clues on the reliability of the extracted information) might give good results regarding the prediction of agreement features for German finite verbs.

6.2 Tense and mood

Register/domain Looking at Figure 3, it becomes obvious that a single English tense can translate into different German tenses. Always choosing the most frequent German tense for a given English tense does not lead to satisfying results (cf. Table 6). On the other hand, Schiehlen (1998), who presented one of the first studies on learning the tense translation from bilingual corpora, stated that this simple tense mapping already achieved the accuracy of 95%. We achieve 70%. This is probably due to register and domain difference: while Schiehlen (1998) worked with corpora related to appointment scheduling (spoken language), we work with news data (written language) which has important differences with respect to tense translation.

Tense usage The correct choice of tense in both human and automatic translation depends on factors which are beyond the scope of our approach (we model the lexical choice of the verbs and syntax). This is true even though some languages have strict tense usage rules. One factor may simply be a *rule* such as the one found in the EC guidelines for translation from English to German[4]: *"Protokolle oder Berichte von Sitzungen werden in der deutschsprachigen Fassung stets im Präsens verfasst..."* / "It is required to use present tense in the translation of protocols and reports, regardless of the tense in the source language." Such a rule does not apply to the translation of news articles. However, in news articles tense/moods are used, in particular subjunctive mood, in which the reporter does not present his own assessment of a situation, but what someone else said (Csipak, 2015), which are almost never used in texts found on the internet (see konjunktiv I + II in Table 3).

Language–pair specific features Ye et al. (2006) presented thoughts about the knowledge that human translators use. The aim was to use this knowledge to model tense translation for Chinese–English. For this specific language pair (and possibly for the corpus used), the knowledge about temporal ordering of the actions was the key information. On the other hand, for English–French, Meyer et al. (2013) found that a *narrativity* feature helps to translate the English past tense into one of the possible French tenses.

Tense switch We observed sentence pairs in which the English is written in past tense, while in German, present tense is used. Obviously, there are contexts in which tense switches are allowed. We assume that these sentences are headlines which allow for this kind of tense variation.

Tense interchangeability It seems that in numerous contexts, tense translation can sometimes even be a matter of taste. Sammon (2002) states that in German the imperfect and perfect are interchangeable in many contexts, the difference between the two tenses being largely stylistic. A similar example is reported speech where Konjunktiv I, Konjunktiv II and indicative tenses are often used interchangeably (Csipak, 2015).

Sequential problem It is also not very clear whether the tense/mood is to be dealt with as a sequential problem or not. On the one hand, in the monolingual context of correcting English tense, Tajiri et al. (2012) argues for a sequential tense model. On the other hand Ye et al. (2006) observed that sequential dependence of the tenses is not as strong as expected. In the bilingual context, there seems to be a strong dependence on the tense in the source language. Statistics about the number of clauses in the sentences shown in Figure 4, shows that our data mostly consists of simple sentences containing only one clause (i.e. one finite verb). In other words, for most of the sentences, an intra-sentence tense sequence is simply not given. Inter-sentence tense modeling, i.e., across sentence boundaries, could be more reasonable, as for example, presented by Gong et al. (2012) for Chinese to English SMT.

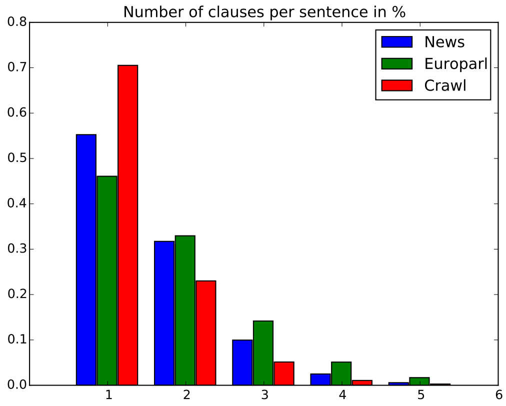

Figure 4: Percentage of sentences from different corpora containing different number of clauses.

Evaluation of the verbs The final question we raise is how to evaluate translations with respect to information related to discourse such as tense and modality (or negation as discussed by Fancellu and Webber (2014)). Automatic evaluation such as BLEU is not appropriate since it compares the translation with the reference mainly on the *lexical* level. What about human evaluation? Our evaluators have a Kappa score of 0.33 which is rather low. The humans thus allow for a certain variance in tense/mood translation which metrics like BLEU cannot capture given only one reference translation. Ideally, we would have multiple references in which all possible tenses are given. Creating such an evaluation test set could be done by gap-filling method proposed by Hardmeier (2014) for evaluation of pronoun translation.

[4] Guidelines for translations into German used by the European Commission: `http://ec.europa.eu/translation/german/guidelines/documents/german_style_guide_en.pdf`

Summary For modeling mood translation, features such as reported speech, conditional context, polite form, etc. would more clearly describe the contexts in which a specific mood occurs. The information about tense ordering proposed by Ye et al. (2006) for Chinese–English would probably be helpful also for English–to–German translation. However, the extraction of such features is more complicated than simply using *surface* features such as words, POS tags, etc.

7 Future work

The verbal inflection handling that we present in this paper is implemented as a post-processing step to the translation. We use the words, i.e. verbs, generated by the SMT system and change them according to our inflection models. An interesting approach would, however, be to use a more abstract representation of German VPs which would allow for generation of all of the words in a VP as specified by the inflection model. For example, we could handle inserting/deleting verbs (auxiliaries), reflexives or even negation.

As for the modeling of tense and mood, we are going to explore possibilities to include discourse knowledge (which was discussed in the previous section) into the classification model. Such a model could also be used within the translation step, for example, to rerank translation alternatives.

Acknowledgments

This work was partially supported by Deutsche Forschungsgemeinschaft grant Models of Morphosyntax for Statistical Machine Translation (Phase 2). This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644402 (HimL) and from the European Research Council (ERC) under grant agreement No. 640550.

References

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-HLT)*, pages 763–770, Columbus, Ohio, June.

Anders Björkelund and Joakim Nivre. 2015. Non-deterministic oracles for unrestricted non-projective transition-based dependency parsing. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 76–86, Bilbao, Spain, July. Association for Computational Linguistics.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, Ann Arbor, Michigan.

Peter Collins and Carmella Hollo. 2010. *English grammar. An introduction*. Palgrave macmillan, 2 edition.

Eva Csipak. 2015. *Free factive subjunctives in German*. Niedersächsische Staats- und Universitätsbibliothek Göttingen.

Federico Fancellu and Bonnie Webber. 2014. Applying the semantics of negation to SMT through n-best list re-ranking. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, April.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling inflection and word-formation in SMT. In *Proceedings of the the European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, France.

Adrià de Gispert and Jose B. Mariño. 2008. On the impact of morphology in English to Spanish statistical MT. *Speech Communication*, 50(11-12):1034–1046.

Anita Gojun and Alexander Fraser. 2012. Determining the placement of German verbs in English-to-German SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 726–735, Avignon, France, April.

Zhengxian Gong, Min Zhang, Chewlim Tan, and Guodong Zhou. 2012. N-gram-based tense models for statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 276–285, Jeju Island, Korea, July.

Christian Hardmeier. 2014. Discourse in statistical machine translation. In *Studia Linguistica Upsaliensia, vol. 14. Acta Universitatis Upsaliensis*, Uppsala, Sweden.

Simone Heinold. 2015. *Tempus, Modus und Aspekt im Deutschen. Ein Studienbuch.* narr studienbcher.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics, July.

Thomas Meyer, Cristina Grisot, and Andrei Popescu-Belis. 2013. Detecting narrativity to improve English to French translation of simple past verbs. In *Proceedings of the 1st DiscoMT Workshop at 51st Annual Meeting of the Association for Computational Linguistics (ACL).*

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA, October. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 2012. DEPFIX: a system for automatic correction of Czech MT outpus. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 362–368, Montréal, Canada, June.

Geoff Sammon. 2002. *Exploring English grammar.* Cornelson Verlag.

Michael Schiehlen. 1998. Learning Tense Translation from Bilingual Corpora. In *COLING-ACL 1998 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 1183–1187, Montral, Quebec, Canada, August.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition, and inflection. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers - Volume 2*, pages 198–202, Jeju Island, Korea, July.

Yang Ye, Victoria Li Fossum, and Steven Abney. 2006. Latent features in automatic tense translation between Chinese and English. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, Sidney, Australia, July.