Automatic Alignment and Annotation Projection for Literary Texts

Uli Steinbach
Department of Computational Linguistics
Heidelberg University

Ines Rehbein
Leibniz ScienceCampus
IDS Mannheim/ Heidelberg University

{steinbach|rehbein}@cl.uni-heidelberg.de

Abstract

This paper presents a modular NLP pipeline for the creation of a parallel literature corpus, followed by annotation transfer from the source to the target language. The test case we use to evaluate our pipeline is the automatic transfer of quote and speaker mention annotations from English to German. We evaluate the different components of the pipeline and discuss challenges specific to literary texts. Our experiments show that after applying a reasonable amount of semi-automatic postprocessing we can obtain high-quality aligned and annotated resources for a new language.

1 Introduction

Recent years have seen an increasing interest in using computational and mixed method approaches for literary studies. A case in point is the analysis of literary characters using social network analysis (Elson et al., 2010; Rydberg-Cox, 2011; Agarwal et al., 2012; Kydros and Anastasiadis, 2014).

While the first networks have been created manually, follow-up studies have tried to automatically extract the information needed to fill the network with life. The manual construction of such networks can yield high quality analyses, however, the amount of time needed for manually extracting the information is huge. The second approach based on automatic information extraction is more adequate for large scale investigations of literary texts. However, due to the difficulty of the task the quality of the resulting network is often seriously hampered. In some studies, the extraction of character information is limited to explicit mentions in the text, and relations between characters in the network are often based on their co-occurence in a predefined text window, missing out on the more interesting but harder-to-get features encoded in the novel.

A more meaningful analysis requires the identification of character entities and their mentions in the text, as well as the attribution of quotes to their respective speakers. Unfortunately, this is not an easy task. Characters in novels are mostly referred to by anaphoric *mentions*, such as personal pronouns or nominal descriptors (e.g. “the old women” or “the hard-headed lawyer”), and these have to be traced back to the respective *entity* to whom they refer, i.e. the speaker.

For English, automatic approaches based on machine learning (Elson and McKeown, 2010; He et al., 2013) or rule-based systems (Muzny et al., 2017) have been developed for this task, and a limited amount of annotated resources already exists. For most other languages, however, such resources are not yet available. To make progress towards the fully automatic identification of speakers and quotes in literary texts, we need more training data. As the fully manual annotation of such resources is time-consuming and costly, we present a method for the automatic transfer of annotations from English to other languages where resources for speaker attribution and quote detection are sparse.

We test our approach for German, making use of publically available literary translations of English novels. We first create a parallel English-German literature corpus and then project existing annotations from English to German. The main contributions of our work are the following:

- We present a modular pipeline for creating parallel literary corpora and for annotation transfer.
- We evaluate the impact of semi-automatic postprocessing on the quality of the different components in our pipeline.
- We show how the choice of translation impacts the quality of the annotation transfer

Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 35–45
Minneapolis, MN, USA, June 7, 2019. ©2019 Association for Computational Linguistics

and present a method for determining the best translation for this task.

2 Related work

Quote detection has been an active field of research, mostly for information extraction from the news domain (Pouliquen et al., 2007; Krestel et al., 2008; Pareti et al., 2013; Pareti, 2015; Scheible et al., 2016). Related work in the context of opinion mining has tried to identify the holders (speakers) and targets of opinions (Choi et al., 2005; Wiegand and Klakow, 2012; Johansson and Moschitti, 2013).

Elson and McKeown (2010) were among the first to propose a supervised machine learning model for quote attribution in literary text. He et al. (2013) extended their supervised approach by including contextual knowledge from unsupervised actor-topic models. Almeida et al. (2014) and Fertmann (2016) combined the task of speaker identification with coreference resolution. Grishina and Stede (2017) test the projection of coreference annotations, a task related to speaker attribution, using multiple source languages. Muzny et al. (2017) improved on previous work on quote and speaker attribution by providing a cleaned-up dataset, the QuoteLi3 corpus, which includes more annotations than the previous datasets. They also present a two-step deterministic sieve model for speaker attribution on the entity level and report a high precision for their approach[1]. This means that we can apply the rule-based sieve model to new text in order to generate more training data for the task at hand. The model, however, only works for English.

To be able to generate annotated data for languages other than English, we develop a pipeline for automatic annotation transfer. This enables us to exploit existing annotations created for English as well as the rule-based system of Muzny et al. (2017). In the paper, we test our approach by projecting the annotations from the English QuoteLi3 corpus to German parallel text. While German is not exactly a low-resourced language,[2] we would like to point out that (i) ML systems can always benefit from more training data, and (ii) that our pipeline can be easily adapted to new languages.

In the next section, we present our approach to annotation transfer of quotes and speaker mentions based on an automatically created parallel corpus, with the aim of creating annotated resources for quote detection and speaker attribution for German literature.

3 Overview of the pipeline

Our pipeline makes use of well-known algorithms for sentence segmentation, sentence alignment and word alignment (figure 1). The entire pipeline is written in Python. Individual components are implemented as classes and integrated into the main class as sub-module imports. The modular architecture facilitates the integration of additional classes or class-methods inside the main class, the replacement of individual components as well as the integration of new languages and more sophisticated post-processing and transfer methods.

Sub-task specific outputs are flushed to file after each step in the pipeline. Thereby, the user is given the opportunity to modify the output at any stage of the process.

3.1 Sentence segmentation

Sentence segmentation is by no means a solved problem (see, e.g., Read et al. (2012) for a thorough evaluation of different segmentation tools). This is especially true when working with literary prose where embedded sentences inside of quotes pose a challenge for sentence boundary detection.

In our pipeline, we use the Stanford CoreNLP (Manning et al., 2014) which offers out-of-the-box tokenisation and sentence splitting. We selected CoreNLP because it offers support for many languages and is robust and easy to integrate. Once the input text is segmented into individual sentences, we need to align each source sentence to one or more sentences in the target text.

3.2 Sentence alignment

Sentence alignment is an active field of research in statistical machine translation (SMT). The task can be described as follows. Given a set of source language sentences and a set of target language sentences, assign corresponding sentences from both sets, where each sentence may be aligned with one sentence, more than one, or no sentence in the target text. It has been shown that one-to-one sentence alignments in literary texts

[1] When optimised for precision, the system obtains a score >95% on the development set from *Pride and Prejudice*.

[2] The DROC corpus (Krug et al., 2018) provides around 2000 manually annotated quotes and annotations for speakers and their mentions in 90 fragments from German literary prose.

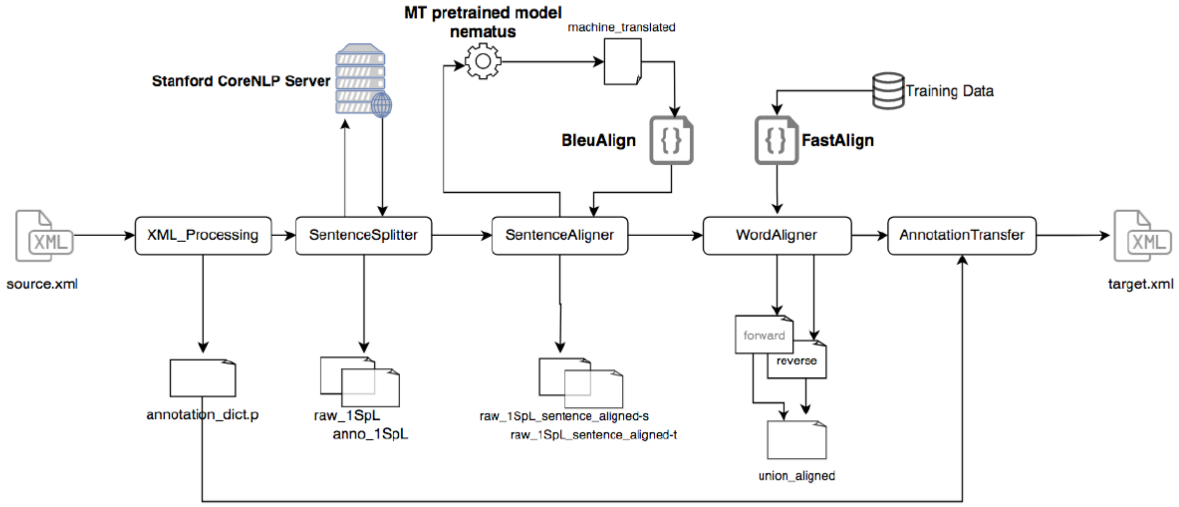

Figure 1: Overview of pipeline architecture and workflow

are less frequent than in other genres (Sennrich and Volk, 2010), and the alignments heavily depend on the lexical choices made by the translator. Even though Manning and Schütze (1999) suggest that, in general, around 90% of sentence alignments are 1:1 alignments, "sometimes translators break up or join sentences, yielding 1:2 or 2:1, and even 1:3 or 3:l sentence alignments" (Manning and Schütze, 1999, p. 468). Sennrich and Volk (2010) manually align a set of 1000 sentences and report only 74% of 1:1 beads, showing that sentence alignments can vary considerably, depending on genre and text type.

While in early days sentence length - measured in tokens or characters - was used as an indicator for parallel text (Gale and Church, 1993a), more recent approaches often use length-based features in combination with lexical similarities for semi-supervised classifier training (Yu et al., 2012; Xu et al., 2015). Mújdricza-Maydt et al. (2013) model sentence alignment as a sequence labelling task and solve it using a CRF sequence classifier.

We use a different approach, proposed by Sennrich and Volk (2010), who first create an automatic translation of the source text, yielding aligned translations for each sentence in the original text. Then, they try to find matching sentences in the automatic translation of the source text and the human-translated target text based on sentence similarity according to the BLEU metric (Papineni et al., 2002).[3]

The alignment itself is based on the computed similarity scores and consists of a two-pass procedure. In the first step, the algorithm is looking for 1-to-1 alignments that maximize the BLEU score for the document, thereby respecting the monotonic order of the sentence pairs. Then, the sentences that remain unaligned are either forming 1:N alignments or are aligned based on a length-based algorithm. Sentences that cannot be aligned in the second pass are discarded.

While the majority of existing tools are not suitable for hard-to-align parallel texts such as literary prose (Sennrich and Volk, 2010, p.1), this approach showed good results on a corpus of historical texts, consisting of yearbooks of the Swiss Alpine Club from 1864-1982. We thus decided to integrate it in our pipeline.

Neural MT with Nematus For translating the source text into the target language, we use Nematus (Sennrich et al., 2017a,b), a neural encoder-decoder model with attention which is similar to Bahdanau et al. (2014).

An encoder (implemented as a bi-directional RNN) reads in word vectors (one vector for each word in a sentence) and generates an output vector of variable length from the sequence of hidden states. Subsequently, the decoder – another bi-directional RNN – learns which words in the source sentence are most relevant for generating a good translation. The model used in this work has been pre-trained with default parameters and configuration (subword segmentation, layer normalisation, a minibatch size of 80, a maximum sen-

[3] BLEU is a standard metric for MT evaluation, based on the overlap of word n-grams in the source and target texts.

tence length of 50 words, word embeddings with 500 dimensions and a hidden layer size of 1024).

Aligning MT and human translation The Bleualign algorithm is composed of two steps. In the first step, the algorithm tries to find a set of anchor points, using BLEU as a similarity score between the machine-translated source text and the human-translated target text. These anchor points are a set of 1:1 alignments considered reliable based on BLEU scores and sentence order.

In a second step, the sentences between these anchor points are either aligned using BLEU-based heuristics or the length-based algorithm of Gale and Church (1993b). The latter algorithm is applied to the target and translated source sentences and functions as a fallback for all gaps with a symmetrical size of unaligned sentences. Sentences that cannot be aligned are discarded.

We use default parameters for Bleualign (a maximum of 3 alternative BLEU-aligned sentences in the first run, a BLEU-scoring restriction on bigrams and second pass gap-filling by means of BLEU and the Gale and Church algorithm).

3.3 Word alignment

Once we have aligned the sentences in our parallel corpus, the next step is the alignment of words between the source and target sentences. We use fast_align (Dyer et al., 2013), a log-linear reparameterisation of IBM Model 2, the second of a set of well-known SMT alignment models developed by IBM in the late 1980s. Fast_align is unsupervised and thus applicable to any language for which training data is available. It outperforms the Giza++ implementation of the IBM Models 1-5 (Och and Ney, 2003) with regard to speed, translation quality (measured in BLEU score) and alignment error rate (Dyer et al., 2013). While the method has recently been outperformed by neural approaches (Legrand et al., 2016), its fast and efficient implementation and decent results make it well-suited for integration in our pipeline.

3.4 Annotation transfer

The final step in our pipeline is the transfer of annotations from the source to the target side. For the task at hand, we directly transfer the speaker and quote annotations based on the word alignments. We hypothesize that this simple and straightforward approach will be sufficient in our case where quotation marks are reliable anchor points for word alignment. Speakers, on the other hand, are often referred to by proper names which, due to string similarity, will also show a high word alignment precision, and we also expect a higher-than-average precision for the alignment of referring noun phrases and personal pronouns.

In the next section, we test our approach and evaluate the individual components of our pipeline for annotation projection from English to German, based on the QuoteLi3 corpus.

	Emma	**P & P**	**total**
quotes	742	1,575	2,317
mentions	399	765	1,164
entities	49	32	81

Table 1: Annotations of quotes, speaker mentions and entities in the QuoteLi3 corpus (*Emma* and *Pride and Prejudice*).

4 Data

For English, the QuoteLi3 corpus (Muzny et al., 2017) provides manual annotations of speakers and quotes in three novels (*Emma* and *Pride and Prejudice* by Jane Austen and *The Steppe* by Anton Chekhov).[4] Since no publically available digital translation for the Chekhov novel was found, our evaluation will focus on the two Austen novels which include more than 2,300 annotations for quotes and more than 1,100 mentions for 81 speakers (table 1).

4.1 Impact of the literary translation

For many novels, not just one but a number of translations are available. We are thus confronted with the problem of having to choose one translation from a set of available texts, and it is not clear how to determine the most adequate translation for the task at hand.

Translation divergences are a known problem for MT (Dorr, 1994; Dorr et al., 2004). In parallel corpora of literary prose, however, divergences are even more prominent than in many other genres. A high-quality literary translation not only needs to transfer the semantic meaning of the source text into the target language but also has to consider stilistic devices such as metaphor, alliteration, hyperbole, oxymoron, simile and more that are difficult to translate. Therefore, the translator often has

[4] The corpus is available for download from https://nlp.stanford.edu/muzny/quoteli.html.

to diverge from the literal translation and resort to a freer phrasing that is more faithful to the underlying meaning or literary function of a certain text passage. This means that different translations of the same text can vary considerably, and the choice of translation for annotation projection might have a crucial impact on the quality of the outcome.

To investigate this issue, we use two different translations for the same novel, *Pride and Prejudice* (PP), in our experiments. The first one is by Karin von Schwab (PP_KS), the second is a translation by Helga Schulz (PP_HS). For *Emma*, a recent translation by Angelika Beck was chosen.

This allows us to evaluate how different translations of the same novel impact the quality of the output for different components in our pipeline.

4.2 Goldstandard

For evaluation, we created two goldstandards, including a total of 600 sentences (300 sentences for sentence alignment, another 300 sentences for word alignment). For each task, we selected 100 sentences from each of the translations (Emma, PP_HS, PP_KS). Sentence selection was not random but focussed on sentences including quotes and speaker mentions. This allowed us to reuse the goldstandard for evaluating the annotation transfer. As a result, sentence length in the goldstandard is slightly higher than the average sentence length in the corpus.[5]

4.3 Settings for evaluation

We compare two different settings in our experiments, (i) a *fully-automatic* setting and (ii) a *semi-automatic* setting. In the *fully-automatic* setting, the texts are extracted from the annotated XML files and directly fed into the pipeline, passing through sentence splitting, tokenisation, MT translation, sentence alignment, word alignment and annotation transfer without any intervention or correction by the user.

In the *semi-automatic* setting, the texts have been subject to a number of genre-dependent pre- and post-processing steps which are described below. These processing steps are adjusted to the text genre and translation specifics and probably need modification and further adaptation when transferred to other literary texts from potentially different domains.

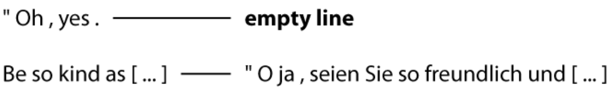

Figure 2: Examples for missing merge in sentence alignment output.

P1: Sentence segmentation Before sentence segmentation, we automatically harmonised punctuation (e.g. “ ” " to ").

After segmentation, incorrectly split sentences were merged again, e.g. splits after short exclamations (*Oh! to be sure*) and after quotes (e.g. *"To be sure!" cried she playfully*). We merged the segmented parts with their preceding or subsequent sentence, based on regular expressions. We also harmonised punctuation (e.g. in the English version, commas are inside quotes while in the German translation, commas were put outside the quote: *“It is one thing,” said she* vs. *“It is one thing”, said she*). These task- and genre-specific processing steps could be done automatically, without manual effort.

P2: Sentence alignment In our experiments, we took empty lines in the output of the sentence aligner as a proxy for alignment errors and manually checked a total of 94 empty lines in the whole corpus[6]. This took – with support of a powerful editor and split screen functionality (Sublime) – less than one hour to complete. Most often, the missing merge was due to divergences in the translation - for example a varying use of punctuation (figure 2).

The impact of the *semi-automatic* pre- and post-processing steps on the quality of the different components in our pipeline are discussed below.

5 Evaluation

5.1 Sentence alignment

As the manual correction of the whole corpus is out of scope for this work, we report three different measures to assess the quality of the sentence alignment module:

1. Recall
2. Comparison against goldstandard
3. BLEU overlap with automatic translation

[5] The avg. sentence length in the goldstandard is 27.4 / 29.5 (Emma / PP), the avg. sentence length for the whole novel is 25.6 / 24.7 / 23.6 (Emma / PP_KS / PP_HS).

[6] Result are heavily dependent on sentence segmentation output, therefore we recommend to implement text- and genre-specific pre- and postprocessing steps for sentence segmentation optimisation.

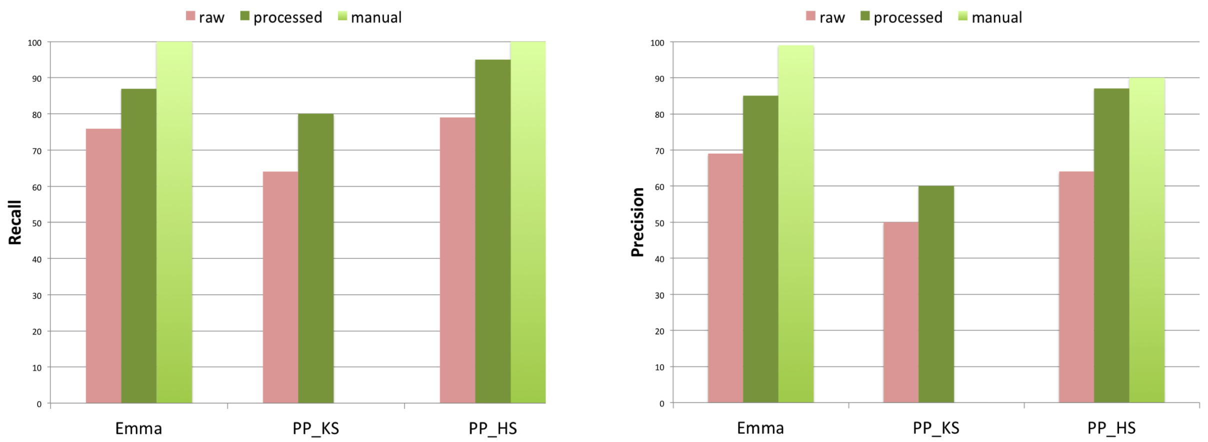

Figure 4: Recall (left) and precision (right) for sentence alignment for different settings (raw: no post-processing; processed: automatic pre-/post-processing; manual: resolution of null-aligned sentences) on the goldstandard.

Recall is computed as the total amount of source sentences in the corpus that have been aligned with (one or more) target sentences. Figure 3 shows that especially for the PP_KS translation, recall in the *fully-automatic* setting is low. However, preprocessing the sentence-segmented XML-input prior to sentence alignment (see P1) can increase recall from below 50% up to 90% and above. For the two other translations, preprocessing results in even higher recall (96% to 100%).

Our second evaluation reports precision and recall on the goldstandard (figure 4). Here we also evaluate the impact of the manual resolution of null-aligned sentences. Both precision and recall for the goldstandard testset increase after automatically pre/post-processing the data. Results show crucial improvements especially for the translation that is closer to the original text (PP_HS). This shows that the selection of the translation has a huge impact on the quality of annotation transfer for literary texts. We also showed that taking empty lines (null alignments) as an indicator for alignment errors can reduce time requirements for manual correction considerably while yielding substantial improvements (precision and recall) for sentence alignment.

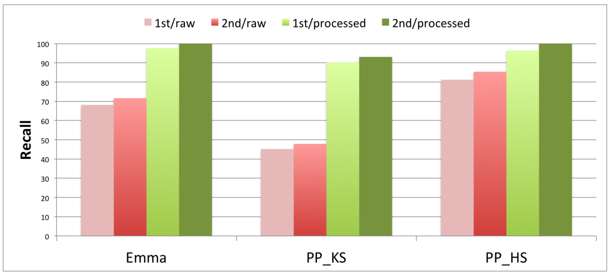

Figure 3: Recall for 1st and 2nd pass of sent. alignment for different settings on the whole corpus (raw: fully-automatic; processed: +automatic preprocessing (P1))

Our third evaluation measure reports the average BLEU (uni- to 4-gram) sentence similarity score between the machine-translated source sentences and their aligned target sentences from the human translations.[7] The automatic translation is expected to be much closer to the original novel than a professional human translation. We can thus take the similarity between the human translation and the *automatic* translation as a proxy for the closeness of the human translation to the *original* novel. We thus hypothesize that the translation of Pride and Prejudice that shows a higher average BLEU similarity to the automatically translated text will be more suitable for annotation projection than a translation with lower similarity scores.

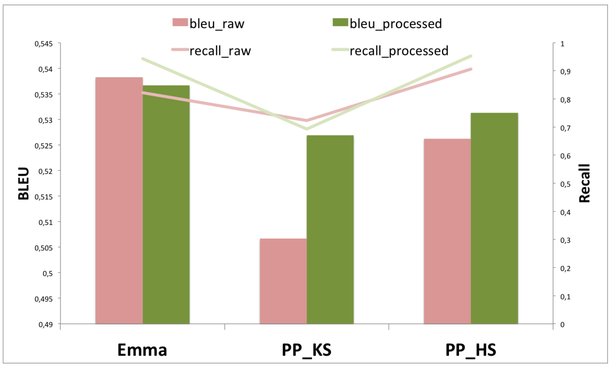

Figure 5: Avg. sentence BLEU score w.r.t source MT (w/wo processing/restricted to 1:1 alignments)

[7] The BLEU scores are calculated for those source sentences that are 1:1 aligned with a target sentence. Recall is thus relative to the amount of first-pass alignments.

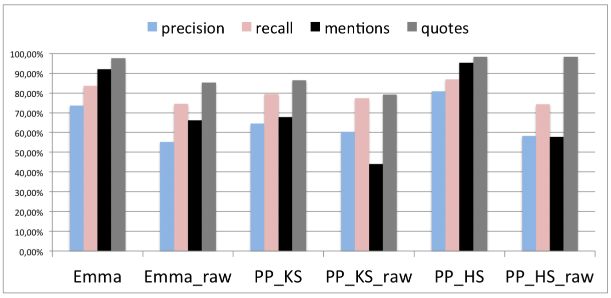

Figure 6: Word alignment evaluation (precision and recall) and precision for transfer of mentions/quotes (goldstandard: all sentences).

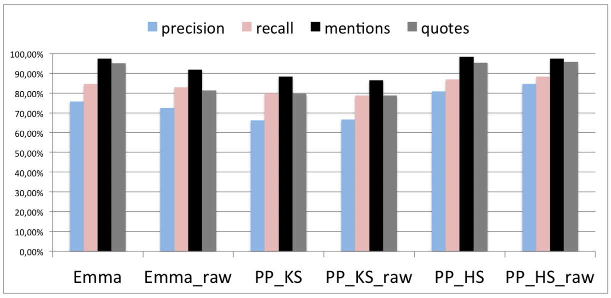

Figure 7: Word alignment evaluation (precision and recall) and precision for transfer of mentions/quotes (goldstandard: correctly aligned sentences only).

Figure 5 shows that BLEU similarity between PP_HS and the MT translation of the source text is much higher than for PP_KS. As expected, BLEU similarity corresponds to a higher recall for sentence alignment, showing that it is indeed a good measure for determining which translation (out of a set of candidate translations) should be chosen for high-quality annotation transfer.

For *Emma*, pre/post-processing did not further increase BLEU similarity, probably due to the already high similarity scores in the raw data. Surprising is the higher recall for PP_KS (raw) compared to the processed data. We can only suspect that due to the low similarity between source and target, alignment quality is low and thus recall on the raw data is unrealistically high and does not reflect the precision of the alignments.

5.2 Word alignment

Word alignment quality depends strongly on the quality of the sentence alignment output. Therefore, we report results for the *fully-automatic* and *semi-automatic* settings. We compare results for all sentences in the goldstandard (figure 6) with the ones we get when evaluating word alignments only on correctly aligned sentences (figure 7). In addition to precision and recall for word alignment (all words), we also report results for a task-based evaluation focussing on the projected annotations for speaker mentions and quotes.

Again, results are substantially higher for the *semi-automatic* setting, showing that our pre/post-processing can prevent error propagation from earlier components downstream. When looking only at those alignments that are relevant for annotation transfer of speaker mentions and quotes, we observe high precision in the nineties. This confirms our hypothesis that direct transfer based on word alignments works well for our task.

As before, we observe significantly higher results for PP_HS, the translation that is closer to the original text than PP_KS. For the transfer of speaker mentions, this increases results from below 70% to around 95%, and for quotes we see an increase from around 87% (PP_KS) to over 98% (PP_HS). The high precision for quote alignments (especially for the *raw* texts) most probably is an artefact of the way quote alignments were evaluated. To count as a true positive, it suffices if the quotation marks are correctly word-aligned to a quotation mark in the source text. This can result in a false positive if the underlying sentences are misaligned, i.e. the quote is incorrectly aligned to a different quote of similar length. Therefore, we also evaluated word alignments on the smaller set of correctly aligned sentences in the goldstandard (figure 7), thus excluding false matches. Here we see a much smaller gap in precision between speaker mentions and quotes, and – naturally – a smaller gap between *fully-automatic* and *semi-automatic* which again emphasizes the importance of error correction in the first stages of the pipeline, especially for sentence alignment.

5.3 Error Analysis

Table 2 shows recall for annotation transfer on the whole dataset. While we observe only a small increase in recall between the *fully-automatic* and the *semi-automatic* setting, please keep in mind that the results do not consider the correctness of the transferred annotations and that recall for the whole dataset should be compared to precision and recall on the smaller goldstandard (figures 6, 7). Below, we present an analysis of the most frequent error types observed on the goldstandard.

Many errors are caused by translation diver-

			PP_KS (raw)	PP_KS (pr.)	PP_HS (raw)	PP_HS (pr.)	Emma (raw)	Emma (pr.)
Quotes found			92,6% (1551)	92,5% (1548)	99,0% (1657)	99,6% (1668)	93,2% (691)	98,8% (732)
of which	**1:1**		66,9% (1038)	69,4% (1074)	83,0% (1376)	87,5% (1459)	76,6% (529)	82,1% (601)
	1:N		23,7% (367)	23,6% (366)	10,4% (172)	9,3% (155)	14,9% (103)	13,5% (99)
	of which	Resolved	55,3% (203)	57,4% (210)	43,0% (74)	27,7% (43)	43,7% (45)	60,6% (60)
		Default	44,7% (164)	42,6% (156)	57,0% (98)	72,3% (112)	56,3% (58)	39,4% (39)
No Alignment			8,7% (146)	6,4% (108)	6,5% (109)	3,2% (54)	8,0% (59)	4,3% (32)
Mentions found			91,9% (751)	92,4% (755)	98,5% (805)	99,9% (816)	92,2% (367)	100 % (398)
of which	**1:1**		60,0% (451)	60,4% (456)	78,1% (629)	83,8% (684)	76,6% (281)	82,2% (327)
	1:N		22,8% (171)	22,6% (171)	13,0% (105)	10,7% (87)	14,7% (54)	15,8% (63)
	of which	resolved	31,0% (53)	31,0% (53)	34,3% (36)	36,8% (32)	50,0% (27)	52,4% (33)
		Default	69,0% (118)	69,0% (118)	65,7% (69)	63,2% (55)	50,0% (27)	47,6% (30)
No alignment			15,8% (129)	15,7% (128)	8,7% (71)	5,5% (45)	8,0% (32)	2,0% (8)

Table 2: Recall for annotation transfer for the whole corpus (raw: fully-automatic, pr.: semi-automatic setting).

gences (figure 8) where the sentence remains partly unaligned. In our example, the content of the English sentence was split into more than one sentence in the German translation. During sentence alignment, however, the German sentence was incorrectly aligned 1:1 to its English pendent. As a result, some of the content is missing, leading to poor word alignment. This type of error needs to be addressed during sentence alignment or in a post-precessing step before word alignment.

The high precision for annotation transfer can be partly explained by the high amount of 1:1 word alignments for speaker mentions and quotes, due to string equality between the word pairs in the source and target texts (e.g. proper names or pronouns for speaker mentions, see table 3).

n-gram	Emma	PP
unigram	254	528
bigram	126	229
trigram	15	7
4-gram	3	1

Table 3: N-gram statistics for mention words (raw frequencies) in the corpus.

A recurring pattern in our data is the incorrect co-alignment of target words to neighbouring tokens, resulting in 1:N word alignments (figure 9). These co-alignments pose a problem for our direct approach to annotation transfer but can be easily resolved using simple string-matching heuristics. As illustration, consider figure 9 where we can simply compare "Lydia" to both alignment candidates on the German side *{Lydia, wollte}* and so identify the correct projection site by string identity.

Unfortunately, this is not always an option. de Marneffe et al. (2009) show that the automatic resolution of multi-word alignments to the right target term is a hard problem and requires automatic recognition of multi-word expressions. For more complex projection tasks, we will thus need a more sophisticated alignment method, based on graph optimisation or machine learning. Previous work in the context of semantic role labelling has followed this approach, with promising results (Padó and Lapata, 2005, 2009; van der Plas et al., 2011; Kozhevnikov and Titov, 2013; Akbik et al., 2015; Akbik and Vollgraf, 2017; Aminian et al., 2017). We would like to explore this further in future work.

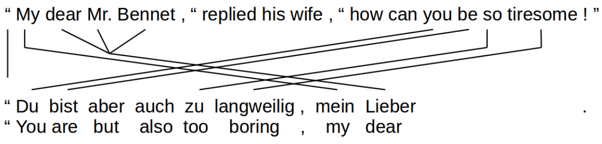
" My dear Mr. Bennet , " replied his wife , " how can you be so tiresome ! "
" Du bist aber auch zu langweilig , mein Lieber .
" You are but also too boring , my dear

Figure 8: Transfer error caused by translation divergence (incorrect 1:1 sentence alignment).

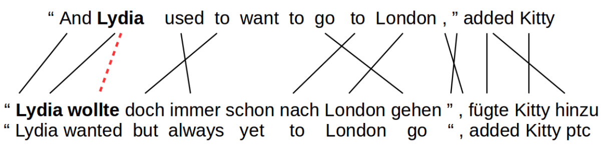
" And Lydia used to want to go to London , " added Kitty
" Lydia wollte doch immer schon nach London gehen " , fügte Kitty hinzu
" Lydia wanted but always yet to London go " , added Kitty ptc

Figure 9: Transfer error caused by incorrect co-alignment.

6 Conclusions and future work

We have presented a modular NLP pipeline for annotation transfer in literary texts.[8] Our pipeline integrates freely available NLP tools into a modular toolkit that allows the user to run the whole pipeline in a fully automatic setting or to perform the different processing steps individually and apply post-processing to improve the quality of the output. The modularity of our toolkit also facilitates the adaptation of individual processing steps and the integration of new components as well as the adaptation to new languages.

Our pipeline can be used for annotation transfer and for the creation of large parallel corpora for computational literary studies, or to bootstrap additional in-domain training data to improve the precision of sentence and word alignment tools for literature.

We identified weak points and possible improvements that we would like to address in future work. One example is the integration of a module (or method) for automatic resolution of multiword alignments after word alignment, or the resolution of null alignments after the sentence alignment step (for example by applying a translation-based sentence similarity measure). Another important issue for future work is to improve annotation projection by replacing the direct transfer based on word alignments with a more sophisticated method based on graph optimisation or ML.

Acknowledgments

This research has been partly conducted within the Leibniz Science Campus "Empirical Linguistics and Computational Modeling", funded by the Leibniz Association under grant no. SAS-2015-IDS-LWC and by the Ministry of Science, Research, and Art (MWK) of the state of Baden-Württemberg.

References

Apoorv Agarwal, Augusto Corvalan, Jacob Jensen, and Owen Rambow. 2012. Social network analysis of Alice in Wonderland. In *Workshop on Computational Linguistics for Literature*, CLfL 2012, pages 88–96.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition banks for multilingual Semantic Role Labeling. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, ACL 2015, pages 397–407.

Alan Akbik and Roland Vollgraf. 2017. The projector: An interactive annotation projection visualization tool. In *The 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2017, pages 43–48.

Mariana S. C. Almeida, Miguel B. Almeida, and André F. T. Martins. 2014. A joint model for quotation attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–48.

Maryam Aminian, Mohammad Sadegh Rasooli, and Mona T. Diab. 2017. Transferring semantic roles using translation and syntactic information. In *The 8th International Joint Conference on Natural Language Processing*, IJCNLP 2017, pages 13–19.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, HLT/EMNLP 2005, pages 355–362.

Bonnie Dorr, Necip Fazil Ayan, and Nizar Habash. 2004. Divergence unraveling for word alignment. *Natural Language Engineering*, 1(1):1–17.

Bonnie J. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20:597–633.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *The North American Chapter of the Association of Computational Linguistics*, NAACL 2013, pages 644–648.

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. 2010. Extracting social networks from literary fiction. In *The 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147.

David K. Elson and Kathleen R. McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *The Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI 2010.

[8] The software will be made available from `https://www.cl.uni-heidelberg.de/research/downloads/` (annot-transfer-lit).

Susanne Fertmann. 2016. Using speaker identification to improve coreference resolution in literary narratives. Master's thesis, Computational Linguistics.

William A. Gale and Kenneth Ward Church. 1993a. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

William A. Gale and Kenneth Ward Church. 1993b. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Yulia Grishina and Manfred Stede. 2017. Multi-source projection of coreference chains: assessing strategies and testing opportunities. In *The 2nd Coreference Resolution Beyond OntoNotes Workshop*, CORBON-2017.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *The 51st Annual Meeting of the Association for Computational Linguistics*, ACL 2013, pages 1312–1320.

Richard Johansson and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3):473–509.

Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *The 51st Annual Meeting of the Association for Computational Linguistics*, ACL 2013, pages 1190–1200.

Ralf Krestel, Sabine Bergler, and René Witte. 2008. Minding the source: Automatic tagging of reported speech in newspaper articles. In *The International Conference on Language Resources and Evaluation*, LREC 2008.

Markus Krug, Frank Puppe, Isabella Reger, Lukas Weimer, Luisa Macharowsky, and Stephan Feldhaus. 2018. *Description of a Corpus of Character References in German Novels – DROC [Deutsches ROman Corpus]. DARIAH-DE Working Papers. Göttingen: DARIAH-DE.*

Dimitrios Kydros and Anastasios Anastasiadis. 2014. Social network analysis in literature. the case of The Great Eastern by A. Embirikos. In *5th European Congress of Modern Greek Studies*.

Joël Legrand, Michael Auli, and Ronan Collobert. 2016. Neural network-based word alignment through score aggregation. In *The First Conference on Machine Translation*, WMT 2016, pages 66–73.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press. Pp. 468.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *The 52nd Annual Meeting of the Association for Computational Linguistics*, ACL 2014, pages 55–60.

Marie-Catherine de Marneffe, Sebastian Padó, and Christopher D. Manning. 2009. Multi-word expressions in textual inference: Much ado about nothing? In *The 2009 Workshop on Applied Textual Inference*, TextInfer 2009, pages 1–9.

Eva Mújdricza-Maydt, Huiqin Körkel-Qu, Stefan Riezler, and Sebastian Padó. 2013. High precision sentence alignment by bootstrapping from wood standard annotations. *Prague Bulletin of Mathematical Linguistics*, 99:5–16.

Grace Muzny, Angel X. Chang, Michael Fang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *The 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2017, pages 460–470.

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

Sebastian Padó and Mirella Lapata. 2005. Cross-lingual bootstrapping of semantic lexicons: The case of FrameNet. In *The National Conference on Artificial Intelligence*, pages 1087–1092.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

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

Silvia Pareti. 2015. *Attribution: a computational approach*. Ph.D. thesis, University of Edinburgh, UK.

Silvia Pareti, Timothy O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *The 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2013, pages 989–999.

Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 299–304.

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news. In *The International Conference on Recent Advances in Natural Language Processing*, RANLP 2007, pages 487–492.

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. 2012. Sentence boundary detection: A long solved problem? In *The 24th International Conference on Computational Linguistics*, COLING 2012, pages 985–994.

Jeff Rydberg-Cox. 2011. Social networks and the language of greek tragedy. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(3):1–11.

Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. Model architectures for quotation detection. In *The 54th Annual Meeting of the Association for Computational Linguistics*, ACL 2016.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017a. The University of Edinburgh's neural MT systems for WMT17. In *The 2nd Conference on Machine Translation*, WMT 2017, pages 389–399.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017b. Nematus: a toolkit for neural machine translation. In *The 15th Conference of the European Chapter of the Association for Computational Linguistics (Software Demonstrations)*, EACL 2017, pages 65–68.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas*, AMTA 2010.

Michael Wiegand and Dietrich Klakow. 2012. Generalization methods for in-domain and cross-domain opinion holder extraction. In *The 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2012, pages 325–335.

Yong Xu, Max Aurélien, and Yvon Francois. 2015. Sentence alignment for literary texts – the state-of-the-art and beyond. *Linguistic Issues in Language Technology – LiLT*, 12(6).

Qian Yu, Max Aurélien, and Yvon Francois. 2012. Revisiting sentence alignment algorithms for alignment visualization and evaluation. In *The 5th Workshop on Building and Using Comparable Corpora*.