Bittersweet Death: German to English Language Machine Translation

Machine Translation is a complicated and complex task. Even for humans fluent in both the source and target languages, translating between them is a non-trivial task. In this paper we explore the difficulties in performing machine translation without the aid of statistical translation methods. For this experiment we chose to translate from German into English. As German and English are closely related languages, the process of machine translation from one to the other would seem to be a task that is not terribly difficult. As we will show in this paper, the similarities of the languages do not preclude significant differences that create significant challenges for machine translation.

1. INTRODUCTION

The German language is a very regular, very structured language governed by the Rat für deutsche Rechtschreibung, a body composed of 39 members from 6 countries who collectively decide the canonical definition of the language. Because of this close governance, the German language has very little room for ambiguity, but the rigid structure of the language can lead to very complex constructions. In a simple German sentence, the subject is in the first position, the verb is in the second position, and the object of the verb is in the third position. The adverbs follow in the order of time, then manner, then place. The subject and object can only appear in the first and third positions in the sentence. The verb must come in the second or last position.

One of the best known traits of the German language is the tendency for verbs to appear at the end of the sentence, often long after the subject and object have been seen or heard. As Mark Twain wrote, ”whenever the literary German dives into a sentence, that is the last you are going to see of him till he emerges on the other side of his Atlantic with his verb in his mouth.” There are five typical reasons that the verb in a German sentence is moved from the second position into the last position in the sentence:

—Dependent clauses — in the case of a dependent clause, the verb is simply moved from it’s regular position into the last position in the sentence. For example, ”I sleep soundly”, translated as ”Ich schlafe tief,” becomes ”..., weil ich tief schlafe,” in the dependent clause, ”..., because I sleep soundly.”

—Past perfect — when a sentence uses the past perfect, the auxiliary verb, typically ”haben” or ”sein”, takes the second position in the sentence, displacing the participle to the end of the sentence. ”Ich schlafe tief,” becomes, ”Ich habe tief geschlafen.”

—Future perfect — with the future perfect tense, as with the past perfect, the auxiliary verb, ”werden”, replaces the verb, sending it to the last position in the sentence as an infinitive. ”Ich schlafe tief,” becomes, ”Ich werde tief schlafen.”

—Modal auxiliaries — when a modal auxiliary is used, it takes the second position in the sentence, moving the verb as an infinitive to the end. ”Ich schlafe tief,” becomes, ”Ich will tief schlafen.”

—Subjunctive/passive — when writing or speaking in the passive voice or subjunctive (Konjuntiv II) tense, the word ”würden” is typically placed in the second position and the infinitive form of the verb is moved to the final position. ”Ich schlafe tief,” becomes, ”Ich würde tief schlafen.”

Taken independently, the above five rules do not present any significant challenge. In the latter four cases, a placeholder verb is left in the second position with which the final verb can be reunited. In the first case, the exact location of the second position has to be rediscovered, but because of the regular rules of the language, locating the second position is always possible, even if occasionally challenging. The above five rules are not limited, however, to independent applications. Several can, in fact, apply to the same clause, compounding their effects. For example, ”Ich schalfe lang, weil ich tief schlafen dürfen werden will,” includes two modal auxiliaries and the future tense in a dependent clause, producing four verbs piled up at the end of the clause.

Other complications with German verbs include the use of separable prefixes, reflexive verbs, a broader set of tenses than in English, and the fact that some inflections are shared among two or more persons, e.g. first person plural, third person plural, and the formal person all share the same verb inflection.

Beyond the verbs, German also presents a challenge in the complexity of the rules governing objects. It is possible for the subject and object of the sentence to be swapped, resulting in a sentence equivalent to, ”him hit I.” In such a sentence, the identity of the subject and object must be discerned by the cases of the pronouns and articles and the inflection of the adjectives. Additionally, a noun in German may be arbitrarily replaced by a definite article which acts somewhat like a pronoun. Combined with the verb reordering caused by dependent clauses and the various verb tenses, these rules can easily result in sentences that are confusing to a human reader and very challenging for machine translation.

One additional challenge in translating between German and English is the mapping of prepositions. While there are some cases where German prepositions map cleanly to their English counterparts, in most cases the correct choice of translation for a preposition is highly dependent on the context in which it is used and is hence difficult to define. Difficulty in mapping prepositions in a common issue in translations and is often addressed through the use of a language model or other statistical method.

As an amusing aside, a common issue in human translations between English and German stems from a class of words known as ”false friends.” False friends are words that are lexicographically similar or even the same in both languages but with divergent meanings. One extreme example is the word ”gift.” In English, a ”gift” is a present. In German, ”Gift” means ”poison.”

2. WORKING CORPUS

We chose a set of 15 sentences from the novel ”Bittersüße Tode” - or ”Bittersweet Death” (translated from the English title ”Guilty Pleasures” – German translators are reknown for changing the titles of the translated works), the first book of a long series of fiction novels by author Laurell K. Hamilton. To select our development set, we searched for 10 sentences that each contain more than one of the sort of challenges outlined above. The test set was collected as the first block of 5 consecutive sentences following the first set of development set sentences. Because they represent a contiguous chunk of text that was selected effectively at random, the test set sentences give a fairly unbiased representation of typical sentences in the German translation of a pulp fiction novel about vampires and wereanimals.

Table I. Development Sentences

1	”Das kurze schwarze Haar hatte er sich aus dem dnnen dreieckigen Gesicht nach hinten geklatscht.”
2	”Er hatte mich schon immer ein wenig an eine Gestalt aus einem Gangsterfilm erinnert.”
3	”Aber fr alle Flle vermied ich es, ihm direkt in die Augen zu sehen.”
4	”Jetzt, wo Willie ein Vampir war, war die Sache mit der Entbehrlichkeit natrlich nicht mehr von Bedeutung.”
5	”Ich wollte ihn fragen, was sich denn nderte.”
6	”Ich wusste nicht, dass Vampire nervse Zuckungen haben knnen.”
7	”Ich schaute den Vampir an, der vor mir sa, und zuckte die Achseln.”
8	”Es steht mir nicht zu, polizeiliche Angelegenheiten mit Ihnen zu besprechen.”
9	”Bert und ich wrden uns ziemlich bald mal unterhalten mssen.”
10	”Er entfernte sich von mir und drehte mir den Rcken zu.”

Table II. Test Sentences

1	”Das war die vernnftigste Vorgehensweise, wenn man es mit Vampiren zu tun hatte.”
2	”Frher war er ein Schleimkbel, jetzt war er ein untoter Schleimkbel.”
3	”Das war eine neue Kategorie fr mich.”
4	”Wir saen in der klimatisierten Stille meines Bros.”
5	”Die himmelblauen Wnde, die Bert, mein Boss, fr beruhigend hielt, machten den Raum kalt.”

3. BASIC TRANSLATION SYSTEM

The basis for our translation system is the direct translation method: for each German word, our system looks up the word in a closed-vocabulary dictionary to find the English translation. Our word translations are taken from the LEO online German-English dictionary (http://dict.leo.org). On top of that basis we applied a series of pre- and post-processing strategies to improve the notably poor results of the blind direct translation. In the following sections we describe and motivate those strategies.

3.1 Language Tagging

Prior to performing any operations on the source sentences, we first process them through the Stanford parts of speech tagger to label all words according to parts of speech. While this strategy has no direct impact on the translation, all of the subsequent strategies other than the language model depend on having tagged words in the sentences.

3.2 Interpolation of Idiomatic Phrases

As with any language, German has a rich set of idiomatic phrases that do not translate literally. In order to provide better translations for these sorts of idiomatic phrases, we implemented a hand-compiled phrase translation table which was applied as a pre-processing step. One challenge in handling German idioms containing verbs is that the verb may be located many words away from the rest of the phrase. This can be seen in the fourth development set sentence in the phrase, ”war... von Bedeutung,” meaning ”mattered.”

While four of our development set sentences contain idiomatic phrases, none of our test sentences contain idioms, so no test set improvement was seen from this strategy.

3.3 Reordering of Verbs in Dependant Clauses

Because of the treatment of verbs in German dependent clauses, translation without reordering the words will often produce unintelligible results. Because the displaced verb may be very far away from it's expected position in the sentence, post-processing reordering techniques are unlikely to be able to restore the correct order. We implemented the reordering of verbs in dependent sentences as a pre-processing step.

Dependent clauses are quite common in German, appearing, for example, in development sentences 3-8 and test sentences 1 and 5. The second phrase of development sentence 6 provides a particularly good example. Translated literally, without reordering, it would read, ”that vampires nervous ticks have can.” With reordering the phrase becomes, ”that vampires can nervous ticks have.”

3.4 Recombining Modal Auxiliary Verb Phrases

Similar to the displacement of verbs in dependent clauses, modal auxiliaries cause verb displacements that can significantly impact the final translation and can be hard to correct in post-processing. We implemented verb reordering for modal auxiliary verb phrases as a pre-processing step.

Modal auxiliaries tend to be common in German sentences, especially in dialogue. Sentences 5, 6, and 9 in the development set all contain modal auxiliaries, though none appear in the test set. To continue the dependent clause example in sentence 6, additionally applying reordering on modal auxiliaries would give, ”that vampires can have nervous ticks.”

3.5 Reorder Past Participle Verb Phrases

Like the other verb displacements discussed above, past participles also can word order translation issues. We implemented the reordering of verbs in past participial phrases as a pre-processing step.

The past participle is typically a very common feature in written German, though less so in conversational German. Only development sentences 1 and 2 contain past participles, which is most likely because the tone of the novel is conversational. In sentence 2, the participle ”erinnert” is displaced by 12 words to end of the sentence, crippling the meaning of a direct translation.

3.6 Rejoin Separable Prefixes

Another common word order difficulty results from words with separable prefixes. A separable prefix is indistinguishable from a regular preposition except by word order, and a separable prefix can significantly alter the meaning of the verb to which it belongs. As with the other verb reordering strategies, we implemented the rejoining of separable prefixes as a pre-processing step.

Sentences 7, 8, and 10 in the development set contain verbs with separable prefixes. Separable prefixes do not appear in the test set. Translating the separable prefix verb from the first clause of sentence 8, ”steht ... zu” directly would give ”stands ... closed.” Applying this strategy would instead yield, ”is entitled to.”

3.7 Subject/Object Ordering

Because subjects and objects can be replaced with arbitrary noun or prepositional phrases, discovering the subject and object of a sentence is difficult. As a pre-processing step we implemented a very simple strategy that looks for a nominative pronoun following a verb and swaps it with everything that precedes the verb.

Even this limited form or subject/object reordering was impactful in our translations. The case of a swapped nominative pronoun appears in the development set in sentences 1 and 3 and in sentence 2 in the test set.

3.8 Verb Tenses

Machine translation systems all have to deal with the issue of translating verb tenses. Extracting the tense of a German verb from its inflection and context and then creating applying that tense to the translated English word is a problem that can be handled by hand-coded rules, but it requires requires a large amount of fairly complicated logic. In this project we did just that, complicated logic included. During translation each source verb is interrogated for tense, and then that tense is created from the English base form. In order to handle irregular verbs correctly, we maintain a look-up table of the 100 most common English irregular verbs (http://www.englisch-hilfen.de/en/grammar/unreg_verben.htm) and about 150 irregular German verbs (http://www.mein-deutschbuch.de/lernen.php?menu_id=21).

This strategy applies to the complete development and test sets and every sentence that includes a verb. The impact on fluency of the resulting translation is significant. A language model could be employed to instead select a tense from a list of possible tenses, but if other sentence details (such as number) are also governed by the language model, then details can be lost. For example, ”the walls make the room cold,” may well be translated instead as, ”the wall makes the room cold,” changing the original meaning. By anchoring the tense to the known source tense, the original meaning is better preserved.

3.9 Language Model for Word Choice

The final strategy we employed was the use of a language model to select for noun number, preposition mapping, and generally improve the translation fluency. In our implementation, each sentence generates a list of all possible combinations of the individual word translations. The language model is then used to select the word choices that produce the most likely sentence as a post-processing step. This general strategy applies to all test sentences and resulted in a non-trivial improvement in the translation was the use of a language model in picking most likely individual word translation. To train our language model, we first replaced all contractions in the source text and then trained a Trigram Stupid-Backoff language model. We tested using two corpora – the first consisting of the original English version of the novel from which we drew our development and test sentences (30,000 words in all), and the second consisting of English versions of 18 of the 24 books in the same series of novels (2.4 million words in all). We used a Trigram Stupid-Backoff language model primarily due to its good performance and ease of implementation. Thus, if the German word ”Film” corresponds to ”movie” or ”coating” in our dictionary, we use a probability score provided by our language model to choose the most likely translation given the sentence in which it appears.

4. COMPARISON WITH GOOGLE TRANSLATE

In this section we compare the reference translation from the original English novel with the translation from Google translate and our own translation.

Book: ”It was standard policy for dealing with vampires.”
Google: ”That was the most sensible way, if you had it to do with vampires.”
Ours: ”That was the most reasonable approach, if you it had with vampire to do.”

Our translation model chose ”vampire” when the correct translation is ”vampires”. One aspect in which Google generally performed better was determining whether nouns are singular or plural. In addition, although our model chose almost identical direct word translations in the second clause as Google's, Google's word rearrangement ultimately leads to a more fluent result.

Book: ”He was a slime bucket, but now he was an undead slime bucket.”
Google: ”Before, he was a slime bucket, now he was an undead slime bucket.”
Ours: ”He was back on slime bucket, he was now on undead slime bucket.”

Although the word rearrangement differs, both translated sentence structures are equally valid. However, our translation makes some questionable direct word translations such as ”on” instead of ”a/an” and ”back” instead of ”before” that lends to a misleading result. Google's translations generally perform better when selecting appropriate prepositions and articles.

Book: ”It was a new category for me.”
Google: ”This was a new category for me.”

Ours: ”This was a new category for me.”

Our results exactly match the Google translation, and both differ from the original sentence by the same word. In this case our translation model performed very well.

Book: ”We sat in the quiet air-conditioned hush of my office.”
Google: ”We sat on the air-conditioned silence of my office.”
Ours: ”We sat in the air-conditioned hush of my office.”

Our translation closely matches the original sentence, differing only by a missing word, which is simply not present in the German translation. In this case, our translation out-performs Google's translation, as the Google translation replaced ”in” with ”on” and chose ”silence” rather than ”hush.”

Book: ”The powder blue walls, which Bert, my boss, thought would be soothing, made the room feel cold.”
Google: ”The azure wall, that Bert, my boss, soothing stopped for, do this area cold.”
Ours: ”The sky-blue walls that kept Bert, my boss, for calm, the room went cold.”

In this case, our translation performed slightly better overall. Our translation has a verb misplacement in the last clause ”the room went cold” which is not present in Google's translation. However, Google's clause ”soothing stopped for” is not even remotely representative of the original clause meaning, ”thought would be soothing”.

5. ERROR ANALYSIS

One does not have to look hard to discover issues with the design of this machine translation system. In this section we'll address two categories of issues: those caused by omissions and those caused by flaws.

5.1 Design Omissions

Given the limited scope of this project, a number of strategies were not implemented for lack of time. The most notable among them are:

- —Noun inflection – the proper inflection for nouns has be determined from context and applied to the translated word.
- —Better subject/object reordering – the strategy we implemented to reorder subjects and objects only works in a limited set of cases. A more complete strategy should be implemented.
- —Adverb reordering – German adverbs always follow the verb in time, manner, place order. Those adverbs could be reordered to produce a more fluent translation.
- —Not – Rather than explicitly translating the particle ”not”, a more refined strategy would negate the verb in a more natural way.

5.2 Design Flaws

Our translation system has two major errors, both of which stem from flaws in the design of the system, and which could likely be fixed by the use of additional statistical and other methods.

The first problem is the issue of verbosity and poor fluency. The translations produced by our system suffered from a mechanical sounding flow, using excessive and in some cases extraneous words because of slavish adherence to the 1:1 direct translation model. These effects are most apparent in the translation of the first test sentence, in which our English translation contains 14 words, one for each German word, whereas the original sentence contains only 8. Because these issues stem directly from the direct translation model, addressing them is challenging. An approach that could be effective is to use the language model not just for word selection, but also for word selective omission. Such a strategy would need to be carefully applied, as removing words from a sentence will most often improve the probability of the sentence as it removes factors from the probability calculation. Rather than working with the probability of the full sentence or clause, it may be effective to work with a fixed number of words on either side of the word being considered for removal. Special treatment would still be required for words on near a sentence or clause boundary.

The second issue is one of word order. In the fifth test sentence, the 1:1 model creates a problem in a different way – it carries the German sentence structure over into the English sentence. Even with the attempted verb reordering, the resulting sentence is choppy and hard to understand. There are subtleties to fluent word ordering that cannot be easily captured by hand written rules. To improve the overall fluency and better preserve the sense of the source sentence, the language model could also be applied to selecting word order. Working in pairs or triplets of words, the highest probability ordering could be calculated. If a significant improvement can be had by rearranging words, then it may be worthwhile. Care must be taken, however to not allow excessive or destructive changes. One could imagine a case where a word could be flipped repeatedly from one end of the sentence or clause to the other, creating a less understandable translation. One possible solution is to limit word swapping to within a clause.

A more subtle issue is that the verb tense strategy implemented in our system is driven by a long list of hand-coded rules. Such a system is necessarily brittle. For example, a verb that only appears in the imperative tense, e.g. ”beware,” would break the existing rules. There are also certain classes of English verbs that will be inflected incorrectly for some tenses, just because of the extreme complexity of trying to handle every possible case. Rather than using hand-coded logic to translate verb tenses, a promising approach would be train a classifier, such as näive Bayes, to classify verb tense and select the watching inflection in the translated word. A difficulty in this approach would be selecting meaningful features. Training the classifier would also require a labeled corpus of verb tenses.

In general, the direct translation model seems to be the primary source of problems in our translations. The most obvious approach to improve the translations across the board would be a wholesale switch to a statistical translation model.

6. CONCLUSION

It is difficult to overstate the level of complexity involved in creating a machine translation system based on a direct translation model. Even translating from a very well defined language like German doesn’t do much to ease the burden. Given our experience on this project, we find it not at all surprising that machine translation did not become commonly accessible until the advent of statistical translation methods.