Technical Reports in Language Technology
Report No 25, 2018
http://www.njas.helsinki.fi/salama

Normalizing English for interlingua

Arvi Hurskainen
Department of World Cultures, Box 59
FIN-00014 University of Helsinki, Finland arvi.hurskainen@helsinki.fi

Abstract

The report demonstrates how English can be normalized to make it more suitable as interlingua in machine translation. Typical to modern English is to omit such critical features as relative pronouns and conjunctions in subordinate clauses. This causes a lot of unnecessary ambiguity, which is difficult to resolve. When English is used as interlingua, such features may be inherited from source language, spelled out in English translation, and then transferred to target language, where they may be necessary to make translation acceptable. Swahili and Finnish require, that such features are overtly spelled out in the language, and they should not be lost in interlingua. The paper uses examples of news test from WMT challenge 2017,

Keywords: interlingua, machine translation, disambiguation.

1 Introduction

It is clear that the global machine translation system should be constructed on the concept of interlingua, a mediating language, which encodes the linguistic information of the source language into such format, that it can be effectively and as unambiguously as possible made use of when translating into target language. A number of suggestions have been made as to which particular existing language could function in this role, or whether a new artificial language should be developed for this purpose.

It seems that English as the practical global lingua franca would be a good candidate for interlingua. It is the most widely used language and the public as well as developers are acquainted with it. If we consider English from the technical viewpoint, it is far from ideal. In fact, the form of English currently used is a badly worn-out version of an ideal language. It has lost most of its morphology as well as the clarity of sentence structure. When word morphology is missing, its consequence is a heavy ambiguity of wordforms. The ambiguity does not pertain only to semantics. It is overwhelming already on such elementary levels as part of speech and syntax. Even the most advanced morphological analyzers of English continue to have big problems in ambiguity resolution.

Although English is not such a language, which we would like to have as interlingua, we do not have a choice. However, we can 'improve' or 'normalize' English, so that it is more suitable as an interlingua in machine translation.

There are two strategies for improving English as interlingua. (a) We can translate from source language (SL) into English so that we use such structures and expressions, which have as little ambiguity as possible. (b) We retain from SL such linguistic information, which might get lost when the expression is translated into English. The first strategy is relevant for all languages. The second strategy depends on the type of the SL, that is, how much and what type of relevant linguistic information it contains. In this paper we concentrate on the first strategy.

2 Missing relative pronouns and conjunctions

The current English tends to omit relative pronouns and conjunctions in subordinate clauses. Obviously, it is assumed that the language user can add them mentally to appropriate places and thus keep control of sentence structure. For a human being this is possible; otherwise such features would not have been lost. For machine translation, and especially for rule-based MT, the omission of such critical features is disastrous and causes a lot of unnecessary rule writing, because in target languages (which usually are less worn-out than English) such features must be expressed.

It is not only the omission of relative pronouns and conjunctions, which causes problems. Also, the choice of alternative words for marking the beginning of a relative or subordinate clause may have tremendous effect on ambiguity. Now, when we use English as an interlingua, we are not at the mercy of the current English writer. We can use the forms, which we deem most suitable for our purposes.

Let us start with the word that, which has three roles. It may be (a) a demonstrative pronoun, or (b) a relative pronoun starting a relative clause, or (c) a conjunction initiating a subordinate clause. In addition, it can be entirely omitted in the roles (b) and (c). The reckless use of this word alone causes immense problems in MT.

Let us see what we can do about it. The first advice is that you should never omit a relative pronoun or a conjunction starting a subordinate clause. The second advice is that you should use unambiguous alternatives if possible. For example, instead of that for relative pronoun, you should use the word which. This has only one other meaning (starting a question), and it is easy to disambiguate.

In addition, the use of comma in the beginning of relative and subordinate clauses helps immensely in disambiguation, no matter how unorthodox this practice sounds nowadays.

3 Normalization of source text and its impact on translation

Below I will demonstrate, what effect the omission of critical words has to translation result. Note that the translation system is not blind to the omission of these words. It tries to insert the missing words in the translation process. Sometimes it succeeds, but at other times it either misses it, or adds words in places, where it should not. This shows that the deletion of critical words in writing causes often almost unsolvable ambiguity.

The examples below are arranged so, that first is the original version of the sentence. After it follows the translation with Salama Translator. Then follows the 'normalized' version of the sentence. After it is the translation, again using Salama Translator.

1.a

We were close to losing that price advantage we got from EU membership.

Me olimme menettämäisillämme tuon hintaedun me saimme EU-jäsenyydestä.

1.b

We were close to losing that price advantage, which we got from EU membership.

Me olimme menettämäisillämme tuon hintaedun, minkä me saimme EU-jäsenyydestä.

Comment:

The added relative pronoun helps to translate correctly.

2.a

Earlier this week 50 officials working among national security wrote an open letter to state that they won't support Donald Trump as a president.

Varhemmin tällä viikolla 50 virkailijaa, jotka toimivat kansallisten turvallisuusvirkailijoiden joukossa kirjoitti avoimen kirjeen todeta, että he eivät tue Donald Trumpia presidenttinä.

2.b

Earlier this week 50 officials, who were working among national security, wrote an open letter stating that they won't support Donald Trump as a president.

Varhemmin tällä viikolla 50 virkailijaa, jotka toimivat kansallisten turvallisuusvirkailijoiden joukossa, kirjoittivat avoimen kirjeen todeten, että he eivät tue Donald Trumpia presidenttinä.

Comment:

The word working does not contain sufficient information for translation. The added information who were working helps to solve the problem, whether it should be translated with present or past tense.

3.a

Rio Olympics is the first sporting event Yle has dedicated a whole channel to.

Rion Olympialaiset on ensimmäinen urheilutapahtuma Yle on omistanut koko kanavan.

3.b

Rio Olympics is the first sporting event, to which Yle has dedicated a whole channel.

Rion Olympialaiset on ensimmäinen urheilutapahtuma, mille Yle on omistanut koko kanavan.

Comment:

Here the relative pronoun is deleted but the preposition to is added to the end to indicate that the relative pronoun is not a subject or object, but rather a modifier. Even this structure could be implemented so that correct translation follows, but this solution is hardly ideal. It is simpler to add the relative pronoun which and move the preposition to in front of it.

4.a

He says that if this growth will not stop, the Government's goal of increasing the employment rate to 72 per cent will be lost.

Hän sanoo, että jos tämä kasvu ei lakkaa, hallituksen tavoite lisätä työllisyysastetta 72 prosenttiin menetetään.

4.b

He says that if this growth will not stop, the Government's goal to increase the employment rate to 72 per cent will be lost.

Hän sanoo, että jos tämä kasvu ei lakkaa, hallituksen tavoite lisätä työllisyysaste 72 prosenttiin menetetään.

Comment:

Here we have an example, where a verb modifies a noun. In the original version it is gerund increasing and in the adjusted version it is infinitive to increase. We note that both versions yield correct translation.

5.a

There were four bombs that exploded in a tourist attraction also favored by the Finns, reports Thai newspaper Bangkok Post.

Oli neljä pommia, jotka räjähtivät matkailuvetonaulalla suosivat myös suomalaiset, raportoi thaimaalainen sanomalehti Bangkok Post.

5.b

There were four bombs that exploded in a tourist attraction, which also was favored by the Finns, reports Thai newspaper Bangkok Post.

Oli neljä pommia, jotka räjähtivät matkailuvetonaulalla, mitä myös suosivat suomalaiset, raportoi thaimaalainen sanomalehti Bangkok Post.

Comment:

Here we have an example of missing relative pronoun combined with passive participial structure. The structure favoured by the Finns could be translated with a participial structure, such as suomalaisten suosima. Such a structure is complex to control, because the agent may have long sets of modifiers, and all these must be moved in front of the verb, which on its part must be converted into adjective. And the agent must inflect according to its role in sentence. This type of structure is easier to implement with a relative structure, as is done in the augmented sentence.

6.a

The quarter responsible for the attacks is not known.

Osapuoli vastuussa hyökkäyksistä ei tiedetä.

6.b

The quarter, which is responsible for the attacks, is not known.

Osapuolta, mikä on vastuussa hyökkäyksistä, ei tiedetä.

Comment:

This example is difficult to translate, because the relative clause does not have a verb. This means, that the writer has omitted the relative pronoun and the finite verb. By adding these missing words which is, the sentence can be translated correctly.

7.a

According to analysts, it is difficult to predict the extent of the effect the people's rage from last spring will have on the autumn's election result.

Analyytikoiden mukaan on vaikeaa ennustaa vaikutuksen määrän ihmisten raivolla viime keväästä olla syksyn vaalitulokseen.

7.b

According to analysts, it is difficult to predict the extent of the effect, which the people's rage from last spring will have on the autumn's election result.

Analyytikoiden mukaan on vaikeaa ennustaa sellaisen vaikutuksen määrää, mitä ihmisten raivolla viime keväästä on syksyn vaalitulokseen.

Comment:

Here we have a particularly problematic case, where not only the relative pronoun is missing. Also the referent of the assumed pronoun is not directly preceding the pronoun. In the structure 'the extent of the effect', the relative pronoun may refer to the word 'extent' or 'effect'. The context does not tell clearly which one is the case. If the referent is 'extent',

translation does not need special measures, because in Finnish genitive structure the possessed comes after the possessor, and it is in this case the last word of the main clause. On the other hand, if the referent is effect, we have to add an extra word sellainen (such) to the main clause to denote, that the referent is not the last but second last word of the main clause. This implementation is taken in this example.

8.a

Trump says he was being sarcastic when claiming Obama was the founder of Isis.

Trump sanoo, että hän oli sarkastinen väittäessään Obama oli Isisin perustaja.

8.b

Trump says that he was being sarcastic when claiming that Obama was the founder of Isis.

Trump sanoo, että hän oli sarkastinen väittäessään, että Obama oli Isisin perustaja.

Comment:

This example contains two subordinate clauses, and both of them have omitted the conjunction. The translation system has managed to insert the first conjunction but missed the second. In the augmented version, the conjunction that has been added, and the translation is correct.

9.a

According to the rescue department, two cars that were driving one after another fell into a ditch, after which the car driving behind collided with the rear end of the car driving in front.

Pelastusosaston mukaan kaksi autoa, jotka ajoivat peräkkäin, putosi ojaan, jonka jälkeen autoa ajamiseen takana törmäsi auton peräpäähän kanssa ajamiseen edessä.

9.b

According to the rescue department, two cars that were driving one after another fell into a ditch, after which the car, which was driving behind, collided with the rear end of the car, which was driving in front.

Pelastusosaston mukaan kaksi autoa, jotka ajoivat peräkkäin, putosi ojaan, jonka jälkeen auto, joka ajoi takana, törmäsi sellaisen auton peräpäähän, joka ajoi edessä.

Comment:

This very complex sentence benefits greatly, if its sections are marked clearly. Also this example contains a structure, where the referent of the relative pronoun in target language is not the last word of the main clause. The auxiliary word sellainen (such) must be added to denote the correct referent.

10.a

Previously many used spears to defend themselves.

Aikaisemmin monet käytetyt keihäät puolustaa itseä.

10.b

Previously many people used spears to defend themselves.

Aikaisemmin monet ihmiset käyttivät keihäitä puolustamaan itseä.

Comment:

In this example the analysis of the sentence easily fails. The word used may be analyzed as adjective and spears as noun, and many as a modifier of the noun. If text contains only full sentences, each sentence contains at least one finite verb. Keeping this in mind, it would be easy to disambiguate this sentence correctly. But now texts contain also headings and titles, which do fulfil the criteria of a sentence. Therefore, it is impossible to disambiguate this sentence. If we add the word people, without affecting the message in any way, we get correct translation.

11.a

The country however was not mentioned among the states one can call directly from the room.

Maata kuitenkaan ei mainittu valtioiden joukossa voi soittaa suoraan huoneesta.

11.b

The country however was not mentioned among the states to which one can call directly from the room.

Maata kuitenkaan ei mainittu niiden valtioiden joukossa, mihin voi soittaa suoraan huoneesta.

Comment:

This example hides the relative pronoun and also the information necessary for correct case of the relative pronoun in TL. By adding the relative pronoun which preceded by the preposition to, it is possible to give precise information for translation. Also here we have the case, where we have to add an extra word ne (gen. niiden, those), because the postposition joukossa is the last word of the main clause in TL.

4 Translation from source language through interlingua into a third language

4.1 Treatment of relative constructions

In this section I will investigate, how the problematic constructions discussed above can be bypassed, when translating from Swahili through interlingua into Finnish. A sample of examples discussed above were first manually translated into Swahili. Then this Swahili text was translated into English using Salama Translator. Then this English text was translated into Finnish using Salama Translator. In fact, translation from Swahili into Finnish is performed in a pipe as a single operation. In order to demonstrate the form of interlingua, below we will see also the intermediate translation.

12.

Trump anasema, kwamba alikuwa sarcastic alipodai kwamba Obama alikuwa mwanzilishi wa Isis.

Trump says, that he was sarcastic when he claimed that Obama was the founder of Isis.

Trump sanoo, että hän oli sarkastinen kun hän väitti, että Obama oli Isisin perustaja.

Comment:

In Swahili, the conjunction kwamba is compulsory. The pronoun is preserved in English translation and carried on into Finnish, where it also is compulsory.

13.

Awali wengi walitumia mikuki kwa kujilinda.

Originally many used spears for protecting themselves.

Alkujaan monet käytetyt keihäät suojella itseä.

Comment:

In this case the problem with the subject is not solved even in Swahili, because also there the word wengi can be used alone as subject. The ambiguity would be solved by using watu wengi (many people).

14.

Kwa mujibu wa wachambuzi ni vigumu kutabiri upana wa athari, ambayo ghadhabu ya watu itakuwa nayo kwa matokeo ya uchaguzi.

According to the analysts it is difficult to forecast the width of the effect, which rage of the people will be with it for the results of the election.

Analyytikoiden mukaan on vaikeaa ennustaa sellaisen vaikutuksen leveyttä, joka ihmisten raivo on sen kanssa vaalin tuloksiin.

Comment:

The compulsory use of the relative pronoun in Swahili makes the translation into English clear, and the sentence structure can be translated correctly into Finnish. The way how

Swahili handles possession in relative clauses becomes transparent. Translation is understandable but not fluent.

15.

Kulikuwa na mabomu manne, ambayo yalilipuka katika senta ya watalii, ambayo pia ilipendwa na Wafini, inaripoti Bankok Post.

There were four bombs, which exploded in the tourist centre, which also was liked by the Finns, reports Bankok Post.

Oli neljä pommia, jotka räjähtivät turistikeskuksessa, mistä myös pitivät suomalaiset, raportoi Bankok Post.

Comment:

Swahili does not have participial structures. Therefore, it is natural to translate relative constructions from Swahili directly into relative structures in English, and then again into relative structures in Finnish, although Finnish could use also participial structures. They are, however, complex to maintain, as was discussed above.

4.2 Treatment of conjunctions in subordinate clauses

English users currently often omit conjunctions, which initiate subordinate clauses. Consider the following extract from news media.

16.

Reid said on Thursday he is predicting Clinton will win the elections.

The sentence contains two subordinate clauses, but they are not marked in any way. If we treat this expression in the same way as above, we get the versions in three languages.

Raid alisema Alhamisi, kwamba anatabiri, kwamba Clinton atashinda katika uchaguzi.

Raid said on Thursday, that he forecasts, that Clinton will win in the election.

Raid sanoi torstaina, että hän ennustaa, että Clinton voittaa vaalissa.

Here is a sentence, where two subordinate clauses are without conjunction.

17.

Humanitarian groups had told the court the shops and restaurants were vital, saying the free meals offered by a state-backed association and other groups did not provide enough food for the growing numbers at the camp.

Here again, three translations of the same sentence demonstrate, how the correct marking of clause boundaries helps in correct translation.

Vikundi vya kihuruma vilikuwa vimesema, kwamba korti na mahoteli zilikuwa muhimu sana, vikisema kwamba milo huru, iliyotolewa na jumyiya za kiserikali na vikundi vingine, havikutoa chakula cha kutosha kwa watu waliozidi katika kambi.

The humanitarian groups had said, that the court and the hotels were very important, when they said that the free meals, which were given by the governmental communities and other groups, did not give food enough to the additional people in the camp.

Ihmisoikeusryhmät olivat sanoneet, että oikeusistuin ja hotellit olivat hyvin tärkeitä, kun he sanoivat, että vapaat ateriat, mitkä antoivat valtiolliset yhteisöt ja muut ryhmät, eivät antaneet ruokaa riittävästi lisäihmisille leirissä.

Comment:

The addition of omitted words in Swahili helps to clarify the sentence structure. The participial structure 'which were given by the governmental communities and other groups' is translated in a straightforward way as mitkä antoivat valtiolliset yhteisöt ja muut ryhmät. A more fluent and natural translation would be mitkä valtiolliset yhtiöt ja muut ryhmät antoivat, but this would require reordering of words, which is prone to errors.

A particularly problematic case is in the sentence below, where the omission of critical words makes the sentence highly ambiguous.

18.

They argued the makeshift shops and restaurants often provided shelter and free meals to those in need.

This could be understood in either of the following ways.

They argued the makeshift shops and restaurants, which often provided shelter and free meals to those in need.

They argued that the makeshift shops and restaurants often provided shelter and free meals to those in need.

Probably the latter interpretation is correct, but the sentence structure does not tell it in any way. Now again, when we start from the Swahili version, we get the following versions.

19.

Walidai kwamba maduka ya muda na mahoteli mara nyingi zilitoa kivuli na chakula bure kwa maskini.

They claimed that the provisional shops and hotels many times provided shade and free food to the poor.

He väittivät, että väliaikaiset kaupat ja hotellit monta kertaa tarjosivat suojaa ja vapaan ruoan köyhille.

Comment:

When missing words are added, the translation from Swahili to Finnish succeeds.

5 Conclusion

In this paper I have demonstrated that by means of normalizing the version of English, which we use as an interlingua in MT, we are able to solve often occurring disambiguation problems. In this process there is nothing artificial or extra work. The source language usually has those critical boundary marks, and they can be transferred into the version of English, which we use as an interlingua. From this version of English, we can then translate the sentence into target language, which most probably requires such features to be explicitly expressed.

Although I have used Swahili and Finnish as test languages, the same method should be applicable to any language.