24.900 Final Squib

Nada AMIN

December 7, 2007

In this paper, I look at machine translation between English and French using the Babel Fish system available at http://babelfish.altavista.com/. I choose to study Babel Fish over Google Translate available at http://translate.google.com. Indeed, Babel Fish uses a rule-based approach to machine translation, while Google Translate applies statistical learning techniques to learn from a huge corpus of translated texts – hence, it makes much more sense to try to figure out some coherent and consistent rules behind a rule-based system like Babel Fish than a statistics-based system like Google Translator, where the reasons for a particular translation depend on gigabytes of impenetrable data.

First, I set out to explore how well Babel Fish performs in cases where English and French differ. My two case studies are (1) compound noun formation (from English to French) and (2) third-person singular pronouns and possessives (from French to English). Both these case studies present a challenge to a machine translation system because they can give rise to a sentence which is ambiguous in the source language but not in the target language, forcing the system to resolve the ambiguity in order to translate the sentence. I will start by presenting these two case studies in turn, giving examples of ambiguous sentences in each. In addition, I will extend the second case study to look at reflexive pronouns. Then, I will explore how Babel Fish deals with ambiguous words that have lexical entries with different parts of speech (for example, ‘desire’ can be both a noun and a verb). Finally, I will try to infer the depth of Babel Fish’s syntactic analysis based on how well it performs, especially in the face of large separations between related constituents.

In English, we can form compound nouns, such as ‘insect eater’ by merging two nouns, with the head noun on the right. In French, such a compound noun gets translated into the head noun, on the left, merged with a PP complement headed by the preposition *‘de’* ‘of’, for example *‘mangeur d’insecte’*, literally ‘eater of insect’. Indeed, Babel Fish translates ‘insect eater’ into ‘*mangeur d’instecte*’. So, we can postulate that when Babel Fish sees two adjacent nouns in English (N_c N_h), it considers the left noun a complement of the right noun, and translates it into (N_h *de* N_c) in French. Now, we can think of ambiguous phrases. In English, we can create a clause that acts like a noun out of a VP using the gerund, for example ‘eating insects’. In French, this would be translated using the infinitive form of the verb, for example ‘*manger des insectes*’. However, some gerund forms, for example ‘learning’ and ‘teaching’, are commonly used nouns. So how does Babel Fish analyze a phrase like ‘learning French’? Is it a compound noun made of two nouns with ‘French’ as the head, in which case it would translate it as ‘*Français d’étude*’ (‘French of learning’)? Or is it a VP that acts like a noun, in which case it would translate it as ‘*apprendre le Français*’ (‘to learn French’)? As it turns out, Babel Fish resolves the ambiguity incorrectly most of the time, translating it into the former. So, ‘learning French is hard’ incorrectly translates to ‘*le Français d’étude est dur*’ (‘French of learning is hard’). Curiously, ‘I love learning French’ correctly translates to ‘*J’aime apprendre le Français*’, showing that Babel Fish somehow takes the surrounding context into account when trying to resolve the ambiguity.

Now, I turn to the second case study, the translation of third-person singular pronouns and possessives from French to English. For nominative pronouns, French doesn’t distinguish between persons and objects (using ‘*il/elle*’ for both) while English does (using ‘he/she’ for persons and ‘it’ for objects). In addition, for the dative pronouns and the possessives, French has one form for both genders (‘*lui*’ and ‘*son/sa*’) while English, for persons, has different forms for each gender (‘him/her’ and ‘his/her’). Hence, when translating a pronoun or possessive from French to English, one has to figure out (1) whether the antecedent is a person or an object (2) sometimes whether the antecedent is male or female (if it’s a person and the pronominal form is dative or possessive). In short, the translation system needs to explicitly guess what some features (whether it’s a person and its gender or whether it’s an object) of the antecedent of a pronominal form is and so implicitly guess what the antecedent is. So this case study allows us to gauge how good Babel Fish is at figuring out (or just guessing) antecedents. I will first list the relevant test cases, and then, infer the rules Babel Fish uses for guessing the antecedents.

1. *Il est venu.* (It came.)
2. *Elle est venue.* (It came.)
3. *Son enfant est venu.* (His/her child came.)
4. *Sa vache est venu.* (Its cow came.)
5. *Il est venu avec son enfant.* (It came with his child.)
6. *Elle est venue avec son enfant.* (It came with her child.)
7. *Il est venu avec sa vache.* (It came with its cow.)
8. *Elle est venue avec sa vache.* (It came with its cow.)
9. *L’homme est venu avec son enfant.* (The man came with his child.)
10. *La femme est venue avec son enfant.* (The woman came with her child.)
11. *La vache est venue avec son enfant.* (The cow came with his/her child.)
12. *C’est venu avec son enfant.* (It came with his/her child.)
13. *L’homme est venu avec sa vache.* (The man came with his cow.)
14. *La femme est venue avec sa vache.* (The woman came with her cow.)
15. *La vache est venue avec sa vache.* (The cow came with its cow.)
16. *Durant ses vacances, l'homme est venu avec sa vache.* (During his holidays, the man came with his cow.)
17. *Durant les vacances de son mari, la femme est venue avec sa vache.* (During the holidays of her husband, the woman came with her cow.)
18. *La femme et sa vache sont venues.* (The woman and her cow came.)
19. *La femme est venue et sa vache aussi.* (The woman came and her cow too.)
20. *L’homme est venu. Il est parti, ensuite.* (The man came. He left, then.)
21. *La femme est venue. Elle est partie, ensuite.* (The woman came. She left, then.)
22. *La vache est venue. Elle est partie, ensuite.* (The cow came. It left, then.)
23. *L’homme est venu. Puis, il est parti.* (The man came. Then, it left.)
24. *La femme est venue. Puis, elle est partie.* (The woman came. Then, it left.)

From these examples, I infer the following rules. When Babel Fish finds an antecedent, it uses the features of this antecedent in the obvious way (is it a person? If so, what is its gender?).

By default, Babel Fish uses the object form ('it', 'its'). When a possessive accompanies certain nouns that must belong to persons (like those denoting family relationships: child, sister, brother, father, mother, relative, etc.), the possessive is translated to 'his/her' by default and the gender of the possessive is fixed by the antecedent only when one is found which is a person or a pronoun with explicit gender marking. Babel Fish's search for antecedent can be pretty robust to sentence structure, as shown by the later examples. My guess is that Babel Fish always consider the subject of a sentence as a possible antecedent when it encounters a pronoun (and the subject almost comes for free as it already needs it to figure out the agreement between subject and verb). However, the last two examples allow us to see the limits of Babel Fish's engine to infer the antecedent: it is confused just by some simple extra linear separation between a pronoun and its antecedent. Hence, I conclude that Babel Fish's search for the antecedent is very weak from one sentence to the next, but pretty robust within one sentence.

Actually, Babel Fish's search for antecedent within one sentence is not so robust in the case of reflexive pronouns and clitics. Though it translates '*L'homme se défend.*' correctly into 'The man defends himself.', Babel Fish translates '*L'homme défend seulement lui-même*' incorrectly into 'The man defends only itself.'. Babel Fish failed to match the pronoun '*lui-même*' with '*homme*'. My guess is that Babel Fish is simply missing the rule that takes into account the antecedent for reflexive pronouns of the second form (*'lui-même'* as opposed to *'se'*). In addition, Babel Fish doesn't do too well with pronominal clitics. First, it doesn't consider them as possible antecedents or as having antecedents:

- '*Sa soeur le défend.*' is translated to 'His/her sister defends it.' disregarding the knowledge of the gender of the object.

- ‘*La soeur de l’homme le défend.*’ is translated to ‘The sister of the man defends it.’ disregarding the possibility that the object is the man.

However, the translations can just be seen as conservative instead of wrong. Indeed, these translations fit with how Babel Fish only considers the subject as a possible antecedent, translating ‘*La soeur de l’homme a pris sa défense.*’ as ‘The sister of the man took <u>her</u> defense.’ (as opposed to ‘<u>his</u>’). Yet, in addition, Babel Fish sometimes omits the clitics entirely in a seemingly inconsistent manner. For example, ‘*Un homme l’a défendu.*’ is translated to ‘A man defended.’ without any pronominal object.

The two previous case studies already highlight that the way to break Babel Fish is to force it to resolve ambiguities. Words themselves can be a source of ambiguities, when they have multiple meanings, even sometimes ranging across parts of speech. For example, these English words have lexical entries in both N and V: ‘desire(s)’, ‘work(s)’, ‘flies’, ‘end(s)’. ‘her’ can be a Det or a N. ‘like’ can be a P or a V. Babel Fish doesn’t do well when confronted with ambiguities, because it doesn’t seem to use any deep analysis to judge whether an interpretation can be valid, even merely syntactically. For example, ‘He never works.’ gets translated into the gibberish ‘*Il jamais travaux.*’ (literally ‘He never works.’ where ‘works’ is a noun). Curiously, Babel Fish correctly translates not only ‘He works.’ into ‘*Il travaille.*’ but also ‘He never works from home.’ into ‘*Il ne travaille jamais de la maison.*’ This shows that Babel Fish gets confused by the separation of the subject and the verb with ‘never’ but it can correct its guess if it is given more information. The gibberish translation of ‘He never works.’ also illustrates that Babel Fish doesn’t seem to favour full sentences over sentence fragments. Another example of that is ‘Flies like blood’ which is translated to ‘*Mouches comme le sang.*’ (literally ‘Flies [are] like blood.’). If Babel Fish was trying to parse ‘Flies like blood’ as a whole sentence, the only possible interpretation would be ‘*Les* *mouches aiment le sang.*' ('Flies like [as in enjoy] blood.'). To its credit, it's the double ambiguity of both 'flies' and 'likes' that seem to confuse Babel Fish as it does fine translating the sentence 'Vampires like blood.' into '*Les vampires aiment le sang.*'.

Finally, I now want to gauge how much deep syntactic analysis Babel Fish exploits by looking at how well it performs in cases were related constituents are linearly separated. For example, '*avoir affaire à*' is a French idiom meaning 'to deal with'. Babel Fish recognizes the idiom when the verb '*avoir*' is right next to '*affaire*' but not when it's separated by a word like '*toujours*' ('always' which comes after the conjugated verb in French because of Verb Raising).

- *L'homme a affaire à la femme.* (The man deals with the woman.)
- *L'homme a toujours affaire à la femme.* (The man always has business with the woman.)

Yet, when '*toujours*' separates the verb from its participle, the idiom is correctly deciphered:

- *L'homme a toujours eu affaire à la femme.* (The man always dealt with the woman.)

However, this last case has no Verb Raising and '*eu affaire*' appears as a linear block. So, Babel Fish fails in the case which exhibits Verb Raising, because the idiom doesn't appear as a linear block. This proves that Babel Fish does not try to recover the deep syntactic structure of a sentence but works with linear blocks instead. In fact, my guess is that Babel Fish deals with Verb Raising in French simply by switching the order of the linear block representing the Adverb with the linear block representing the Verb. For some idioms, the translation is done before the switching (like for '*avoir affaire à*') while for others, the translation is done after the switching (like for '*avoir envie de*'). This would explain why '*Tu as toujours envie de venir.*' correctly translates to 'You always want to come.'. In addition, '*Tu as toujours vraiment terriblement envie de venir.*' correctly translates to 'You always really terribly want to come.', which could be explained by postulating that 'always really terribly' is treated as a linear block. We can assess the shallowness of Babel Fish's syntactic analysis by combining idioms with conjunctions of coordination. In particular, '*Tu as absolument et terriblement envie de venir.*' is literally translated to 'You have absolutely and terribly desire of coming': since no switch was performed, the idiom is lost, and the sentence barely English. Again, '*Tu as absolutement et terriblement eu envie de venir.*' is correctly translated to 'You absolutely and terribly wanted to come.' because the idiom 'eu envie de' appears as a linear block. Incidentally, this last example shows us that Babel Fish treats composed conjugation ('*as eu*') differently from just another idiom to be translated after switching. My guess is that Babel Fish only looks at the auxiliary verb to figure out the composed tense but ignores it otherwise. Under this hypothesis, Babel Fish would be robust to errors in choice and agreement of auxiliary verbs, and indeed, it is.

In conclusion, Babel Fish performs relatively well in translations between French and English, because the grammars of these two languages are relatively similar. However, their differences sometimes lead to ambiguities which must be resolved during translation. As I have shown, Babel Fish is not very good at resolving ambiguities, which, therefore, are a good way to break the system.