Pragmatics: Machine translation with Systran

Preben Wik and Anna Hjalmarsson

Introduction

We have evaluated Systran (http://www.systransoft.com/index.html). The language pair English to Swedish was chosen because these were languages we were both familiar with. The three texts used in the evaluation are taken from the course literature on machine translation (Hutchins, 2003 & Mitkov and Barbu, 2002). There were a lot of inaccurate translations and it was not always easy to tell what the cause of the incorrect translation was. For some of the categories we found it more interesting to include translations where Systran had actually managed to translate correctly.

1. Nominal compounds

Systran makes attempts to merge noun phrases with multiple nouns into Swedish compounds. Systran successfully translates “machine translation” into “maskinöversättning”. Despite the fact that Systran often correctly tries to merge compounds “machine translation” was one of the few occurrences of actually correctly translated compound we could find. Not even compounds which are likely to appear in a dictionary such as “sea horse” (havshäst) are correctly translated. Systran’s efforts to join together noun phrases into Swedish compounds are also done when a part of the compound is out of vocabulary. These compounds still appear to be semantically comprehensible to a human:

```
(1) anaphora resolution -> anaphoraupplösning
```

Some of the compounds can probably not be found in a Swedish dictionary but are quite innovative:

```
(2) collocation preference -> kollokationpreferens
```

Many compounds are incorrect because Systran tend to classify verbs as nouns and nouns as verbs:

```
(3) tended to isolate movie palaces -> ansat till
    isolatmoviepalaces
(4) Co-occurrence patterns -> Co-händelsen mönstrar
```

2. Proper names

There is apparently a set of rules in Systran for handling proper names. It is reasonable to assume a rule that would treat all words spelled with capital initial, that are not in the dictionary as a proper name (e.g. Nasukawa). It is not obvious however what to do with words that also have a meaning, such as many surnames have (e.g. Mr Black)

Trying to see how the rules are applied in Systran we created some test sentences.

```
(5) Time will tell how MT systems will improve ->
    Time skar berättar hur MTsystem skar förbättrar
```

Changing case from /Time/ to /time/ makes:

```
(6) tid skar berättar hur MTsystem skar förbättrar.
```

Time, written with capital letters is thus assumed to be a proper name. Looking further into the matter, this does not seem to be consistent however.

```
(7) Old Bridge Street -> gammalt överbryggar streeten
```

There seem to be several rules involved, giving unpredictable results. One would for example expect the proper names Black, White, and Green, to be treated equally, but they are not.

```
(8) I read in the New York Times yesterday that Mr. Black won
    the Prophet championships ->
    Jag läste in New York Times igår att Mr. Svärta segrade
    Prophetmästerskapen.
```

Changing /New York Times / to /Times/, removing the /Mr/ and changing /Prophet/ to /White/ yields:

```
(9) I read in the Times yesterday that Black won the White
    championships->
    Jag läste in tiderna igår att svarten segrade de vita
    mästerskapen.
```

Just changing /Black/ to /White/ resulted in yet another output:

```
(10) Jag läste in tiderna, igår som den White segrade de vita
     mästerskapen
```

And changing to /Black/ to /Green/ became:

```
(11) Jag läste in tiderna igår att gräsplan segrade de vita
     mästerskapen.
```

In the following example a nominal compound rule is influencing the result.

```
(12) Blackforest cake -> Blackforest tårta
(13) Black Forest cake ->Svart skogtårta
(14) Black-Forest cake ->Svärta-Skog tårta
```

3. Tense choice

Systran has not made many mistakes related to tense choice in our translations. Some of the problems that actually do occur are due to Systran’s inability to inflect the Swedish ”ska” (will).

```
(15) will safely -> skar säkert
(16) would be to exploit -> skade är att exploatera
```

Another occurrence of an incorrectly translated tense choice is:

```
(17) as some leaders hoped -> som några leaders hoppas
```

4. Anaphora pronouns

We could only find one occurrence of an incorrect anaphora pronoun in our translations:

```
(18)Whether a system is inherently unimprovable or whether it
    is capable of improvement –>
    Huruvida är ett system naturligt unimprovable eller
    huruvida är den kapabel av förbättring
```

“Den” refers back to system and should consequently be translated as “det”.

5. Inappropriate retention of source language structures

Translations with inappropriate retention of source language structures are fairly common in our translations. Many of them result in word order mistakes and incorrect prepositions. Here are some examples where the English language structure is incorrectly retained into the Swedish translation:

```
(19)than in the days of -> än i dagarna av
(20)once a day per person -> en gång en dag per person
(21)what better time -> vilken mer väl tid
```

6. Mistakes in morphology

Morphology is very often incorrect. Nouns are inaccurately inflected, adjectives do not agree with nouns and verb formations are wrong. Here are two examples where the translations result in a correct morphology:

```
(22)offices  -> kontor
(23)the red planet -> den röda planeten
```

According to the translations of nouns in our texts it seems like Systran contains rules for regular Swedish nouns. Still, the wrong rules are often applied and many nouns are incorrectly inflected:

```
(24)organisations -> organisationar
(25)brilliant inventors -> briljant uppfinnarearna
```

Moreover, adjectives do not always agree with their nouns.

```
(26)last ten years -> sist tio åren
(27)upside down ice cream cones –> uppochnervänd glasskottar
(28)damaged -> skadada
```

And verb formations are wrong:

```
(29)hit -> slågit
(30)would, will -> skar, skade
```

Systran also tries to inflect English words, which are out of vocabulary, using Swedish rules.

```
(31)the Russian leader -> den ryska leaderen
(32)Nobel Peace Prize -> Nobel peacen som var pris
```

7. Word order mistakes

Word order mistakes are fairly common. This category is related to "Inappropriate retention of source language structures" since the incorrect word order is caused by retention of the English language structure:

```
(33)  whether a system is -> huruvida är ett system
```

The following example illustrates that Systran makes some attempts to correct for word order. The position of "unfortunately" (tyvärr) is shifted. Unfortunately the sentence contains several errors and is difficult to comprehend.

```
(34)but unfortunately information on general co-occurrence
    patterns may not be widely available -> men information
    på allmän co-händelse mönstrar tyvärr kan inte vara brett
    tillgängligt
```

Another example of successful correction for word order is:

```
(35)you have found -> har du funnit
```

8. Prepositions

Prepositions are likely to be a troublesome aspect of MT. Language pairs such as English-Swedish are however likely to be less difficult that some other language pairs, because they share so much in structure.

```
(36)whether it is capable of improvement -> huruvida är den
    kapabel av förbättring

(37)systems of the "first generation" -> system av "den
    första utvecklingen"

(38)man of the year -> man av året

(39)My husband, Ed-> Den min maken, Ed
```

9. Verb forms

There are examples of incorrect verb forms. However, most of them are related to morphology; i.e. plural nouns are not considered etc. Another problem is translations of auxiliary verbs. Unfortunately Systran is not capable of inflecting "will" (ska) correctly.

```
(40)time will tell -> time ska berättar
(41)will perform -> skar utför
```

10. Phrases and common compounds

One aspect of MT where one could imagine an improvement in the translations at a relatively low cost, is in the addition of a lexicon of standard phrases and common compounds. We have seen examples of phrases that are translated correctly, such as:

```
(42)in many respects -> I mångt och mycket
```

Indicating the presence of a ’phrase library’ in Systran. (or is it done by a stochastic process?) The use of such a library is not unproblematic either, as the next sentence can give an indication of.

```
(43)after an early morning-> efter en otta
```

Here it would have been better to translate word by word, instead of using “otta”. We have however seen numerous places where such a library of phrases and common compounds would have been helpful, such as:

(44) pine trees-> sörja trees
(45) ice cream cones-> glasskottar
(46) Nobel Peace Prize-> Nobel peacen

It is interesting to see how certain phrases that are translated in a ‘phrase-like’ manner are still wrongly translated.

```
(47)On the other hand->å ena sedan
```

How is this possible? Is it the effect of a statistic process? Or has it simply been edited by a sloppy person? A double check, translating Swedish-English shows that.

```
(48)Å ena sidan-> on the one hand
(49)Å andra sidan-> on the other hand
```

Whereas translating English-Swedish

```
(50)on the one hand->Å ena sidan
(51)on the other hand-> Å ena sidan
```

11. Verbalization of nouns

A common mistake made by Systran is misinterpretation of nouns as verbs. It would be interesting to use part of speech tagging on the source text and use the tags to disambiguate words which can be both verbs and nouns in English.

```
(52)on the ground -> på det slipat
(53)pine trees -> sörja trees
(54)light on this question -> inget ljus på denna
    ifrågasätter
(55)a research paper presents -> pappers- gåvor
```

12. General difficulties with any multi-clause sentence

It is not easy to find a multi-clause sentence without translation errors. To point out, from the multitude of errors in such sentences, which errors that stems from the fact that it is a multi-clause sentence, is however hard. Below is an example of a sentence that seems to have fallen apart in the translation. It is hard to find the structure in the Swedish translation, and to assign values to the different clauses.

```
(56) Finally, it is plain that more and more large
     organisations can and do save translation costs by using
     MT systems as aids in the production of draft
     translations.->
     Slutligen är det vanligt att mer och större
     organisationar kan, och att göra räddningöversättning
     kostar, genom att använda MTsystem som, bistår i
     production av formulerar översättningar.
```

Conclusions

We have looked at some of the linguistic phenomena enumerated in Hutchins' paper, and also how the system handles nominal compounds, proper names, and tense choice. We have also observed a few other phenomena that frequently created errors in the Systran system. Many errors where English nouns were translated into Swedish Verbs could have been resolved by a part-of-speech tagger. Many common words, such as 'tree' and 'morning' were not in the dictionary. Other words were in the dictionary but not in their most common form (e.g. letter-> märka). The system was not capable of resolving genitive.

Our general impression of the quality of translations was not very favorable. Perhaps Systran are not focusing so much on the Swedish part, and have put more focus on getting good translations in other, bigger language pairs. From the example paragraphs we have given in this rapport, it is hard to see how anyone can save translation costs by using such an MT system as an aid in the production of draft translations.

Translations

Source text I

A question frequently asked by those new to the field is whether machine translation has improved – in the last five years, or in the last ten years, or in the last twenty, etc.
In many respects, the answer is quite easy. It is obvious that programs for automatic translation run much faster than in the past because computers are faster. Equally obvious is that text input is much easier than in the days of punched cards and paper tape, and that the output is much easier to read than in the days of almost illegible computer printouts all in upper case. MT systems are also becoming cheaper by the year, even by the month: what cost over two thousand dollars five years ago is now on sale for under five hundred dollars. Finally, it is plain that more and more large organisations can and do save translation costs by using MT systems as aids in the production of draft translations. Twenty years ago this would have seemed almost inconceivable.

Translation I

En ifrågasätta som frågas vanligt av de som är nya till fältet, är huruvida maskinöversättningen har förbättrat - i de sist fem åren, eller i de sist tio åren eller i de sist tjugona, Etc.en I mångt och mycket är svaret ganska lätt. Det är tydligt att program för automatiskt översättningen kör mycket snabbare än i förflutnan, därför att datorer är snabbare. Lika tydligt är att textförlaga är mycket lättare än i dagarna av stansade kort och pappers-tejpa, och det efterbehandlingen är mycket lättare att läsa än i dagarna av nästan oläsligt datorall utskrift i bokstav. MTsystem är också passande mer billig vid året even vid monthen: vad kostar över två tusen dollar fem år sedan, är nu på rea för under fem hundra dollar. Slutligen är det vanligt att mer och större organisationar kan, och att göra räddningöversättning kostar, genom att använda MTsystem som, bistår i production av formulerar översättningar. Tjugo år sedan som detta skade, har verkat nästan inconceivable.

Source text II

This paper has not been concerned with the improvability of any particular system – or any particular type of system architecture. Whether a system is inherently unimprovable or whether it is capable of improvement is a complex question involving system architecture, rule bases, dictionary entries, etc. In the 1980s it was often claimed that systems of the “first generation” with ‘direct translation’ architecture were inherently limited in potential translation quality, and that the later “second generation” systems of the ‘transfer’ type were more likely to produce good quality output. This investigation can throw no light on this question – which is now, in any case, overtaken by the reorientation of MT research to data-driven and corpusbased methods. It should be noted, however, that advocates of some of these new methods now make similar claims about the inherent limitations of rule-based methods and the superiority of statistics-based and example-based approaches. Time will tell whether systems based on these newer approaches will perform better than those which have been investigated here.

Translation II

Detta pappers- har inte angå med improvabilityen av något särskilt system - eller någon särskild typ av systemarkitektur. Huruvida är ett system naturligt unimprovable eller huruvida är den kapabel av förbättring är ett komplex ifrågasätter härska baser, ordboktillträden, Etc., på åttiotalet som den var, gälla systemarkitektur ofta fordrat att system av ”den första utvecklingen” med ”riktar översättning”, arkitektur var naturligt inskränkt i den kvalitets- potentiella översättningen och det det mer sistnämnd var system ”för den andra utvecklingen” av ”den överförings” typen mer rimlig till den goda jordbruksprodukter kvalitets- efterbehandling. Denna utredning kan kasta inget ljust på denna ifrågasätter - som är nu, i något fall overtaken av reorientationen av MT forska till datadrivet och corpusbased mrthods. Det bör noteras, emellertid, att förkämpar av några av dessa som är nya metoder gör nu liknande fordrar om de naturliga begränsningarna av rule-based metoder och superiorityen av statistik-baserat och exempel-baserat att närma sig. Time ska berättar system som baseras på dessa nyare, att närma sig huruvida skar utför mer väl än de som har utforskats här.

Source text III

Co-occurrence patterns act as an important factor (collocation preference) in anaphora resolution which have been successfully exploited in several projects (Dagan and Itai 1990, 1991; Mitkov 1998), but unfortunately information on general co-occurrence patterns may not be widely available. An alternative would be to exploit collocations within the document(s) under consideration (Nasukawa 1994), but due to the scarcity of patterns, this preference cannot frequently be made use of (Mitkov 1999). On the other hand, access to subcategorisation knowledge (selectional restrictions) could be even more difficult. Without such types of information, the majority of anaphora resolution approaches would have problems with examples such as

Translation III

Co-händelsen mönstrar agerar som en viktig faktor (kollokationpreferens) i anaphoraupplösning vilka har lyckat exploaterats i flera, projekterar (Dagan och Itai 1990, 1991; Mitkov 1998), men information på allmän co-händelse mönstrar tyvärr kan inte vara brett tillgängligt. Ett alternativ skade är att exploatera kollokationer inom dokumenten under övervägande (Nasukawa 1994), men tack vare mönstrar scarcityen av, denna preferens kan inte vanligt görs bruk av (Mitkov 1999). Ta fram till subcategorisationkunskap, å ena sidan (selectional begränsningar) kunde vara även svårare. Utan sådan typer av information, majoriteten av anaphoraupplösning att närma sig skade har problem med exempel liksom.

Sentences

1. Did you ever wonder why pine trees look like upside down ice cream cones?

   Undrade du någonsin varför sörja trees ser lika uppochnervänd glasskottar?

2. One of the principal purposes of the Cabinet is to advise the President on any subject he may require relating to the duties of their respective offices.

   En av det främsta ämnar av det kabinett är att råda presidenten på några betvingar honom kan kräva att förbinda till arbetsuppgiftarna av deras respektive kontor.

3. How much car insurance is enough for a totally broke law student?

   Hur mycket bilförsäkring är nog för en totalt pank lagdeltagare?

4. Clouds are classified into a system that uses Latin words to describe the appearance of clouds as seen by an observer on the ground.

   Moln klassificeras in i ett system som använder latin uttrycker för att beskriva det utseendemässigt av moln som sett av en observatör på det slipat.

5. Most studies of America's movie palaces have been nostalgic, preservation-oriented efforts which have tended to isolate movie palaces in time and space from other public architecture and from the larger current of consumerism in the U.S.

   Mest studies av Amerika moviepalaces har varit nostalgiska preservation-orienterade försök som har ansat till isolatmoviepalaces i tid och utrymme från annan offentlig arkitektur och från den större strömmen av consumerism i U.S.

6. In the past, Microsoft has limited its detailed comments to the monthly bulletins, responding to other issues with short statements.

   I förflutnan har Microsoft begränsat dess specificerade kommentarer till de månatliga informationerna som reagerar till annan, utfärdar med kort meddelanden.

7. The hard problem of consciousness is how to explain a state of consciousness in terms of its neurological basis.

   Det hårda problemet av medvetenheten är hur man förklarar ett statligt av medvetenheten benämner in av dess neurological bas.

8. King's renown continued to grow as he became Time magazine's Man of the Year in 1963 and the recipient of the Nobel Peace Prize in 1964.

   Konungrenown fortsatte för att växa, som han blev den Time tidskriftens man av året i 1963 och mottagaren av den Nobel peacen som var pris- i 1964.

9. A research paper presents the results of your investigations on a selected topic.

   Pappers- gåvor för en forska resultaten av dina utredningar på ett utvalt ämne.

10. Unlike some famous persons, he cannot be summed up by a few remarkable quotations.

    I motsats till några berömda personer kan inte han summeds upp av anmärkningsvärda quotations för fåtal.

11. Desperate attempts to save a 76-year-old Liverpool man’s life after an early morning fire at a home on Old Bridge Street on Tuesday morning were unsuccessful.

    Desperat försök till räddningen 76 som årigt Liverpool manliv, efter en otta har avfyrat på ett hem på gammalt överbryggar streeten på tisdagmorning, var mislyckade.

12. Vandalism has hit a new low in Airdrie.

    Vandalism har slågit en ny low i Airdrie.

13. Different characters in the show assist in developing different themes.

    Olika tecken i showen hjälper, i framkallning av olika themes.

14. There is still more to learn from one of the most brilliant inventors and entrepreneurs of modern times, a man who shaped not only industrial America, but also mass entertainment and contemporary culture with his breakthroughs in sound recordings and motion pictures.

    Det finns stilla mer att lära från en av de mest briljant uppfinnarearna och entreprenörerna av moderna tider, en man, som formade inte endast industriella Amerika, men samlas också underhållning, och samtida kultur med hans genombrott låter inspelningar och vinkar in föreställer.

15. Confidence is high that the probe will safely slip into orbit around the red planet.

    Förtroende är kicken som sondera skar säkert snedsteget in i orbit runt om den röda planeten.

16. Forty-eight house trailers and 20 automobiles were damaged or destroyed.

    Forty-eight hussläp och 20 bilar var skadada eller förstörda.

17. If your goal is finding happiness, you have found the right place

    Om ditt mål finner lycka, har du funnit rätten förlägger

18. the spirit is high but the flesh is weak

    anden är kicken, men köttet är svagt

19. My husband, Ed, was not interested in spiritual things and since I am not a forceful person, I did not get involved in a church

    Den min maken, Ed, intresserades inte i andlig saker, och, sedan jag inte är en kraftfull person, fick jag inte involverad i en kyrka

20. Our mission is to boldly go where no other coffee Web site has gone before.

    Vår bestämmelse är fett att gå var ingen annan kaffewebbplats har väck för.

21. Our out-of-this-world site contains information about every aspect of coffee, and our intent is to elevate appreciation and awareness of quality coffee globally.

    Vår ut-av-denna-22#världen plats innehåller information om varje aspekt av kaffe, och vårt uppsåt är att höja gillande och medvetenhet av kvalitets- kaffe globalt.

22. Technological innovations seem to trickle down and find their way into our lives either directly or indirectly

    Technological innovations verkar för att sippra besegrar och finner deras långt in i våra liv endera direkt eller indirekt.

23. The Russian leader expressed no contrition for Soviet domination of Eastern and Central Europe, as some leaders hoped

    Den ryska leaderen uttryckte ingen ånger för sovjetisk dominans av östliga och centrala Europa, som några leaders hoppas

24. Like a strange cross between object and plant, cactuses appealed to Takashi Murakami because of their almost Martian appearance.

Gilla ett konstigt korsar anmärker och planterar between, kaktus som nästan appelleras till Takashi Murakami på grund av deras utseendemässiga Martian.

25. What better time to try your hand at creative writing than on a nice warm spring day.

    Vilken mer väl tid till det ditt försök räcka på idérik writing än på en trevlig varm fjäderdag.

26. Exercise is a proven and fun way to reduce stress and improve your overall conditioning

    Öva är bevisat och ett gyckel långt att förminska spänning och förbättra ditt total-villkora

27. When you click the "Give Free Food" button (once a day per person) at http://www.thehungersite.com, this simple action gives over a cup of fortified food to a hungry person.

    Ge fri mat” knäppas (en gång en dag per person) när du klickar ”, på http://www.thehungersite.com, som denna enkla handling ger över en kupa av stärkt mat till en hungrig person.

28. I read in the New York Times yesterday that Mr. Black won the Prophet championships

    Jag läste in New York Times igår att Mr. Svärta segrade Prophetmästerskapen.

References

Hutchins, J. 2003. "Has Machine Translation improved?" An expanded version [PDF, 288KB] of a paper presented at MT Summit IX: *Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA, September 23-27, 2003, 181-188. [East Stroudsburg, PA: AMTA.] [PDF, 191KB]

Mitkov, R. and Barbu, C. 2002. "Using corpora to improve pronoun resolution." *Languages in context*, 4(1). (pdf)
http://clg.wlv.ac.uk/papers/mitkov02.pdf