Patent documentation - comparison of two MT strategies

Lene Offersgaard Claus Povlsen

Center for Sprogteknologi
University of Copenhagen
Njalsgade 80, DK-2300 Copenhagen S
Denmark
{loff,claus}@cst.dk

Abstract

This paper focuses on two matters: A comparison of how two different MT strategies manage translating the text type of patent documentation and a survey of what is needed to transform a MT research prototype system to a translation application for patent texts. The two MT strategies is represented by PaTrans - a transfer and rule based system being used for more than 15 years by the translation agency Lingtech A/S and SpaTrans - a SMT system based on the Pharaoh framework. The SMT systems are characterised by shorter development time and low development cost compared to rulebased systems.

The distinctive text type of patents pose special demands for machine translation and these aspects are discussed based on linguistic observations with focus on the users point of view. Two main demands are automatic pre processing of the documents and implementation of a module which in a flexible and user-friendly manner offers the opportunity to extend the lexical coverage of the system. These demands and the comparison of the two MT strategies are discussed on the basis of proofread patents.

Introduction

Due to the characteristic features of patent documentation, this text type constitutes specific challenges for machine translation. This paper gives a brief description of patent documentation and how well two different MT systems are able to meet these challenges.

The first section gives an introductory description of the text type, patent documentation. Section two and three contain descriptions of the MT systems, a rule-based and an SMT-based system, respectively. The following sections introduce the evaluation procedure and report on the evaluation made on the two MT systems' translational results. The next section goes through the various error types that can be identified in the translational results. Some concluding remarks are given in the next section, summing up the observations that have been made with respect to comparing the translational results generated by the two MT systems. The final section outlines future plans on how to improve the translation quality of the SMT system.

Patent documentation – text typical feature

Since patent documents are official and juridical documents, they are kept in a departmental style meeting the following criteria:

- try to be as factual and impartial as possible
- let all information of given topic be expressed within one period.

The first criterion forms part of the reason why patent documentation texts have proven suitable for automatic translation. The demand of factual language usage promotes occurrences of many non-ambiguous technical terms. In addition, only the concrete and denotative meaning of words from the general word register are used. Even though patent texts are characterized by the absence of polysemiotic readings of the words used (facilitating the MT task), the whole idea or rationale behind writing a patent application makes certain demands that have to be met. The introduction of inventions lead per definition to coining of subject specific terms, designating the new concept in question. With respect to lexical coverage within the area of patent documentation the ratio of new terms will per definition be disproportionately high regardless of the size of the already system known terms.

In other words, an important design requirement of an MT-system tailored to patent documents is that it is capable of – one way or the other – treating system unknown words in flexible and user friendly way. Otherwise it would often result in poor translation results.

The second criterion entails the occurrence of very long sentences with many embedded subclauses and series of prepositional phrases. Again, in order to achieve high quality machine translation results the MT systems must be designed in such a way that treatment of very long sentences does not involve a profound decrease in translation quality. Another general feature embedded in the patent documentation text type is the frequent occurrences of entities such as references to other patents, dates, measure units and text internal references.

While the above mentioned characteristics cover patent documentation in general, other elements in domain subsets of patent documentation - related to the problem of system unknown terms - require specific treatment.

Focus in this context will be on the domain specific area of Chemistry. Not surprisingly this subset of patent documentation is dominated by the presence of many chemical formulae. The syntax of how chemical substances can be combined is well defined though they can be very complex, cf. the following examples:

$-CH_2CH_2N(R^{15})CH_2CH_2$
N-[3-[4-(6-fluor-1,2-benzisoxazol-3-yl)-1-piperidinyl]propyl]phthalimid
2-(3-(2-(ethoxy)ethylcarbonyloxy)propyl)ethyl

In some specific cases chemical formulae are language specific and need to be translated, but in general the formulae are language neutral and can be transferred/translated directly to the target language in question.

It is, however, crucial that MT systems in their design have encompassed a procedure for treating these non-verbal entities in order to obtain a reasonably high translational quality.

Comparison of two MT strategies

In the following a comparison of the use and performance of a statistical phrase-based MT system and a traditional rule-based MT system is made. The comparison focuses on linguistic aspects in the different kinds of error types which were identified in the output of the SMT system. First, we briefly describe the two systems to illustrate the very different nature of the two systems.

The PaTrans MT system

PaTrans is a rule based MT system designed for English-Danish translation of patent texts. PaTrans is a transfer based system directly descended from the Eurotra MT research prototype (EUROTRA, 1991). The transition from research prototype to a production MT system included extensions for optimisation, syntactic error recovery, grammatical coverage of patent document specific phenomena, integration of a part-of-speech tagger, document handling (with preservation of layout information), a rule based entity recogniser and implementation of an automatic post-editing tool (see Ørsnes et al, 1996; Povlsen & Bech, 2001 for a more detailed description).

In addition, in order to facilitate the manually conducted pre-editing task, various tools have been implemented, i.a. a term-coding tool and a tool that by making lookups in the existing term databases can identify system unknown words/terms in the source document.

The SpaTrans MT system

The SpaTrans system is developed in a research project financed by the Danish Research Council. The research concerned evaluation of the feasibility of developing SMT for the Danish language. The focus was on translation of patents from English to Danish. Two patent translation companies participated in the project acting the role of a potential future user and evaluated the potential of SpaTrans system.

The SpaTrans system is based on the phrasal SMT decoder Pharaoh (Koehn et al., 2003; Koehn 2004). The Pharaoh decoder is the translation engine and is placed in surroundings of pre and post processing components. The pre and post processing components are much simpler than the corresponding PaTrans components, but handle some of the same challenges, though they leave others unsolved for the time being. The possibility of using terminology databases and preservation of input document layout are not yet implemented in the SpaTrans system, while preservation of special characters, tokenisation and casing are handled. The SpaTrans system is based on a phrase table and a language model. The Pharaoh training software is used to train the phrase table. The training corpus consists of translated and sentence aligned patents. Experiments using europarl languages training material in combination with the patent texts lead to poorer results on development test set, so it was decided to do the training based only on patent texts at this stage. A similar observation is done by (Simard, 2007). The training corpus size can be seen in table 1. The training resulted in a phrase table with 2.3 mill phrases.

Corpus	English words	Danish words
Training	4.2 mill	4.5 mill
Language model	-	4.5 mill
Devel. test	19.464	17.465
Test	10.035	10.574

Table 1: Sizes for training and test corpus

The sentence length in the training material for Danish sentences is 25 words and for English sentences 28 words. The treatment of formulae and figures are not as elaborated in the SpaTrans system as in the PaTrans system, but regular expression substitutions are performed to solve the most widely used conversion problems between English and Danish figures and references.

The language model is trained (order 3) using srilm (Stolcke, 2002) based on the Danish part of the patent training corpus. Experiments based on human evaluation have shown that the use of the monotonic translation option is best suited for English-Danish translation. We are well aware that the quality of the translations by the SpaTrans system might improve if more training material could be added, but the issue here is mainly to investigate the potential in the use of SMT in Patent translations using domain text resources and to point out strengths and weaknesses. One important limitation of the SpaTrans system is that no terminology database is used. The input format of the decoder allows for applying information about predetermined translations of single words and multiword units. This facility can be used to apply specific terminology to the translation engine and before bringing the system in production use, this will be added to the pre processing module.

Evaluation

Analyses of the output of the two systems are based on BLEU metric (Papineni et al., 2002). There is much focus on evaluation of SMT and MT-systems and the used BLEU metric is only one simple way to measure quality. For a brief overview of other currently used evaluation metrics used for SMT and MT and recent experiences within the field, see Callison-Burch, 2007.

The BLEU metric gives one score for each test document. It has been argued that an increase/decrease in the value of the BLEU score does not guarantee a better/worse translation quality (Callison-Burch et al., 2006). But nevertheless the metric is widely used to measure development improvements in systems.

Given one or more reference translations the BLEU metric is normally used to score a text or a larger test corpus. The BLEU metric can also be used to calculate a score for each sentence in a test corpus, and these sentence based scores are in our evaluation used to focus on sentences with a low score, excluding very short sentences which by the definition of the algorithm will have a low score. Another aspect of the evaluation is the reference translation which is a product of post editing the PaTrans output by an experienced proofreader. This gives a large advantage to the PaTrans system, and this is reflected in the BLEU scores of the test material of the two systems, see table 2.

BLEU	Test patent A	Test patent B
PaTrans	**0.539**	**0.610**
SpaTrans reord.	0.439	0.399
SpaTrans mono.	**0.448**	**0.501**
Diff (PaTrans - SpaTrans mono.)	0.091	0.111

Table 2: BLEU scores for two test documents. Test patent A consists of 227 sentences and test patent B consists of 376 sentences.

Evaluation – one step further

About SpaTrans in general

A general observation concerning SMT systems is that the corpus used as training data per definition reflects the translation performance of the SMT system. As training data are collected within a specific text type about a domain specific subject, will in some cases involve that the SMT system suggests translations that are too narrow in their scope leading to poor evaluation results.

To give a translation example from the SpaTrans system[1]:

Further, these paints, properly formulated and applied, have the ability to remain effective for 5 years.

Has been translated into:
Yderligere, disse malinger, korrekt formuleret og påført, har evnen til at forblive der er effektiv i 5 år.

Literally translation:
Further, these paints, properly formulated and applied, have ability_the to to remain which is effective for 5 years.

The translation of 'effective' to 'der er effektiv' appears at first glance to be somewhat odd and it seems surprising that SpaTrans chose that translation. The Pharoah platform gives access to the word lattice generated during the translation process containing a list of the n-best translations that the system has considered. By adding an additional parameter in the command line, i.e. '–lattice' two files are generated. One that contains the word lattice and another that gives additional information about the states in the word lattice.

Opening the first file and looking up the n-best translations of 'effective', gives the following information:

```
(19638 (22478 "effektiv"          0.0117515))
(19638 (22485 "der er effektiv"   0.00482768))
(19638 (22472 "effektive ,"       0.000421076))
(19638 (22469 ", der effektivt"   0.00057635))
(19638 (22470 "er effektivt"      0.000124405))
(19638 (22471 "effektive med"     7.80866e-05))
-      -      -      -      -      -
-      -      -      -      -      -
```

The first number, 19638 refers to the particular state i.e. the token in the input sentence that is to be translated. The number 19638 contains the word coverage vector of 11111111111111100000, considering the transition probabilities between state 15 and 16 in the input sentence (i.e. between 'remain' and 'effective' in the source sentence). The number in the second column links to the translation of the next token in the input sentence.

As can be seen the best scores for translation of effective are the one in bold, i.e. *effektiv* with the score 0.0117515 and *der er effektiv* with the score 0.00482768. Seen from this isolated point of view it seems as if the model would select the translation 'effektiv' instead of 'der er effektiv'. If you, however, go through all the states and multiply all the probability transitions involved, it turns out that the best path (the least cost demanding path) is the one that has 'effective' translated as 'der er effektiv'. A quick look into the training data confirms that in patent documentation within the chemical subject domain, this translation will be the right one in most cases.

Reordering

Based on contrastive knowledge about English and Danish and various experiments conducted, it was decided that the parameter reordering value was set to the value of '-monotone', i.e. no reordering (see table 2). In terms of word order English and Danish are quite similar. One difference, however, can be seen in sentences in which adverbials (adverbs and prepositional phrases) have been topicalised, i.e. occurring in the first position of a sentence. While you in English preserve the SVO order, Danish swifts into a VSO order. In addition, since many adverbials in Danish cannot occur in position 1 of a sentence, you have to make a reorder to get the syntactically correct position translating from English into Danish. To give an example:

Indeed, marine antifouling paints based on organotin acrylate polymers have dominated the market for over 20 years.

The SpaTrans output:
Faktisk, marinbegroningshindrende malinger baseret på organotin- acrylatpolymerer har domineret markedet for over 20 år.

[1] It should be born in mind that the evaluation reference texts are the results of post-edited outputs from the PaTrans system which without any doubt in comparison with the Spatrans system favours PaTrans.

The post-edited version:
Begroningshæmmende skibsmalinger baseret på organotinacrylatpolymerer har faktisk domineret markedet i mere end 20 år.

Literally translated:
Marine antifouling paints based on organotin acrylate polymers have indeed dominated the market for over 20 years.

When sentences with topicalised adverbials are not reordered the resulting word order in Danish will be muddled-up. This is punished in the BLEU-evaluation and it contributes to the explanation of why the overall SpaTrans BLEU-scores are lower than the corresponding PaTrans BLEU scores.

Agreement

Another error pattern in the SpaTrans translation results is the frequent occurrence of agreement errors, such as in:

This constant erosion of the paint ...
Dette konstante erosion af maling ..

In Danish the noun, 'erosion' has the gender masculine and since the determiner 'dette' has the gender neuter, the translation has an agreement error. Seen from a BLEU score point of view these agreement errors are not crucial. Bearing in mind, however, that these errors are extremely frequent it helps explaining why PaTrans performs better than SpaTrans.

About PaTrans

Although the PaTrans system produces output of a quite high quality (reference to the BLEU-scores), some defective translation results are unavoidable especially in connection with automatic translation of patent documentation. The very high average sentence length requires the implementation of various robustness features ensuring that the system always produces a translation of an input sentence of whatever length. Whenever this failsoft component of the system takes over, it leads to activation of a more lean linguistic analysis of the input sentence which again leads to less precise translation results. The loss of information of morpho-syntax in these cases, for instance, results often in either mistaken or non-inflected word target translations.

Some concluding remarks

The core engines of SpaTrans and PaTrans perform approximately equal. If the problems concerning fronted adverbials and the agreement discrepancies mentioned above were solved then it would be likely that the two systems in terms of translation quality would perform approximately equal. This conclusion illustrates excellently the advantages of the SMT strategy. If you have access to parallel corpora of a high quality, it is possible to develop an SMT-system fast and at low cost that in terms of translation quality performs quite well.

As mentioned above, the step from the Eurotra research prototype to the PaTrans production system required both extensions and improvements of the system. These changes were made in order to tailor the system to process domain specific documents adhering to patent documentation text type.

In this context it would be relevant to call attention to two important PaTrans extensions. First a few comments about the implementation of the automatic entity recogniser. Patent documents contain per definition many entities of generic nature (chemical formulae, patent references etc.) which – seen from an SMT point of view - would require an almost infinite amount of training data to be included in the coverage of the system. Based on this assumption, it would be necessary to implement a preprocessor the functionality of which would be to identify these entities and mark them up so that the SMT system by handling these entities systematically can preserve its translation quality level.

Patent documentation contains per se new concepts and terms leading to a disproportionately high rate of system unknown words. In PaTrans these circumstances have been met by implementing an unknown word detection facility and in addition a user friendly term coding tool. Using SMT systems translating patent documentation would also require some kind of facility (e.g. a user dictionary) that would enable the user to extend the lexical coverage with system unknown terms before the translation process is activated. The need for a user dictionary has been recognized by the big SMT vendor Language Weaver since they have made it possible for the users to add existing term based and dictionaries to the phrase-tables. Language Weaver, however, point to the fact on their website, that in the world of statistical translation adding a user term base to the system could cause some disruption, since the language translation software is based on the probabilistic integrity of the phrase table. Language Weaver recommends alternatively that the user extends the SMT system coverage by including representative texts (containing the user coded terms) in the parallel corpus whenever an extension of the lexical coverage is needed.

SMT systems such as SpaTrans provide good translation results at low costs if good and many parallel data are available. Using SpaTrans in a commercial production context translating patent documentation would require that the functionality of the system is extended with an automatic entity recogniser and that the user of the system – one way or the other – is given the possibility of changing and extending the system lexical coverage in a flexible way.

Future work

In order to improve the BLEU-scores of SpaTrans the agreement problem reported above will be investigated. The experiments will follow two paths.

One will try to find out whether the general assumption – all other factors being equal – that more training data will improve the SMT outcome, as suggested in (Simard 2007). In this experiment both subject domain specific data and data from a general language corpus will be included, training both a general and a domain phrase table and combining these.

The other method will be to enrich the language model with additional linguistically based knowledge. This experiment will be conducted by tagging all the words in the corpus (which the language model is based on) with morpho-syntactic knowledge, by computing probability scores for sequences of these morpho-syntactic tags and finally by integrating these scores in the language model. This experiment will be made within the Moses framework[2]. At the workshop we will present the results from these two experiments.

Acknowledgements

The work reported here was partly financed by the Danish Research Council. We would like to thank Lingtech A/S and Plougmann & Vingtoft for providing us with training material and proofread patents. We would also like to thank the other participants in the SDMT-SMV project.

References

Callison-Burch, Chris and Fordyce, Cameron and Koehn, Philipp and Monz, Christof and Schroeder, Josh, "(Meta-) Evaluation of Machine Translation" in *Proceedings of the Second Workshop on Statistical Machine Translation*, June, 2007, Prague, Czech Republic, Association for Computational Linguistics, pp. 136-158.

EUROTRA (1991). Copeland, C., Durand, J., Krauwer, S. & Maegaard, B. (Eds.). S*tudies in Machine Translation and Natural Language Processing*, Vols. 1 and 2. Luxembourg: CEC.

Koehn, Philip Och, Franz and Marcu Daniel. Statistical phrase-based translation. In Procedings of the Human Language Technology and North American Association for Computational Linguistics Conference (pp. 127—133). Edmonton, Canada. Association for Computational Linguistics, 2003.

Koehn, Philip. Pharaoh: A beam-search decoder for phrase-based statistical machine translation models. In Procedings of AMTA, 2004.

Koehn, Philip, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello Bertoldi, Nicola Cowan, Brooke Shen, Wade, Moran, Christine, Zens, Richard Dyer, Chris Bojar, Ondrej Constantin, Alexandra and Herbst, Evan. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic, June 2007.

Maegaard, B. & Hansen, V. (1995). PaTrans: Machine Translation of Patent Texts, from Research to Practical Applications. In Engineering Proceedings of the Second Language Convention (pp.1—8).

Papineni, Kishore, Roukos, Salim, Ward, Todd and Zhu Wei-Jing. BLEU: A method for automatic evaluation of machine translation. In Procedings of ACL, 2002.

Povlsen, Claus, Bech A. Ape: Reducing the Monkey Business in Post-Editing by Automating the Task Intelligently. In Proceedings of MT Summit VIII (pp. 283—286). 2001, Santiago de Compostela, Spain.

Simard, Michel, Cyril Goutte & Pierre Isabelle. Statistical Phrase-based Post-editing. In Proceedings of NAACL HLT 2007 (pp. 508-515). Association for Computational Linguistics (ACL), 2007.

Stolcke, Andreas. SRILM - An Extensible Language Modeling Toolkit. In Proceedings of Intl. Conf. Spoken Language Processing, Denver, Colorado, September 2002.

Ørsnes, B., B. Music and B. Maegaard. PaTrans – A Patent Translation System. In Proceedings of COLING 1996 (pp. 1115-1118). Copenhagen. 1996.

[2] Moses is a open source drop-in replacement for the Pharaoh decoder. As a new facility Moses offers the possibility of using factored translation models. Factored translation models can be built based on surface forms, lemmas, part-of-speech and morphology.