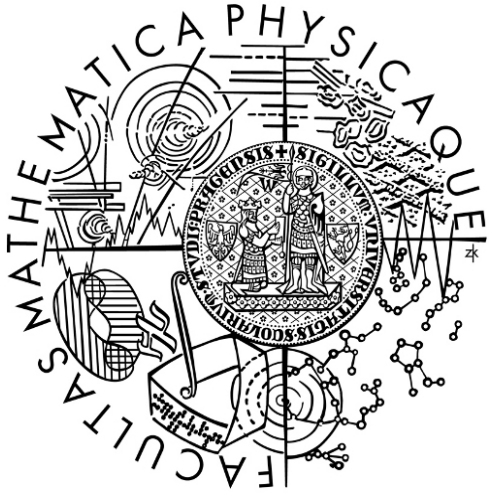
Charles University in Prague
Faculty of Mathematics and Physics

Master Thesis



Fahim A. Salim

Combining Outputs from Machine Translation Systems

Institute of Formal and Applied Linguistics

Supervisor: Ing. Zdenek Zabokrtsky, Ph.D.

Study program: European Masters Program in Language and Communication Technology (LCT)

Charles University in Prague

Prague, 2010

Acknowledgments

First of all I would like to thanks Allah the almighty, the creator of everything for all the blessings, and then to my parents for their unconditional love.

I would also like to thanks my supervisor Dr Zabokrtsky for his guidance at every step of the work, and to LCT coordinators in Prague Dr Lopatkova and Dr Kubon for their support and guidance throughout the year.

I certify that this master thesis is all my own work, and that I used only cited literature. I agree with making this thesis publicly available.

Prague, October 27, 2010 Fahim A. Salim

Abstract

Due to the massive ongoing research there are many paradigms of Machine Translation systems with diverse characteristics. Even systems designed on the same paradigm might perform differently in different scenarios depending upon their training data used and other design decisions made. All Machine Translation Systems have their strengths and weaknesses and often weakness of one MT system is the strength of the other. No single approach or system seems to always perform best, therefore combining different approaches or systems i.e. creating systems of Hybrid nature, to capitalize on their strengths and minimizing their weaknesses in an ongoing trend in Machine Translation research.

But even Systems of Hybrid nature has limitations and they also tend to perform differently in different scenarios. Thanks to the World Wide Web and open source, nowadays one can have access to many different and diverse Machine Translation systems therefore it is practical to have techniques which could combine the translation of different MT systems and produce a translation which is better than any of the individual systems. Since output combination is an additional step over actual translation, therefore it should be very resource and time efficient to be practically usable, and these techniques should also work only on individual system output without bothering much about how the translation was generated because such information is usually not available.

This thesis investigates system output combination techniques. The focus is on techniques which are individual system and language pair independent and are efficient enough to be usable in variety of application scenarios.

Table of Contents

List of Tables

1 Introduction

This work is related to Machine Translation which is a subfield of Computational Linguistics.

1.1 Machine Translation

Machine Translation MT or sometimes referred as Automated Translation is the branch of Computational Linguistics which deals with usage of software for translation from one natural language to another. Material to be translated can be in either speech i.e. spoken language, or it can be in written form i.e. text.

There are thousands of natural languages spoken in the world, with the concept of globalization and with the inventions of newer and faster communication means people from different cultures speaking different languages are coming closer and interacting with each other more than ever before. This cross language or culture interaction introduces the need to translate between languages, and with this interaction comes the need to make available, the knowledge and information present in one language into another. E.g. a scientific paper written in Chinese shared with Czech or Urdu speakers.

The requirement of Translation of all sort of material ranging from ordinary conversations to legal and literary texts on a massive and continuous level seems impossible to be fulfill by human translators alone e.g. a person speaking English per say cannot always acquire a human translator for the language of the country he visits. Organizations like UN, EU or multinational corporations spend fortunes on armies of translators of different languages to run their operations but there is always room for more.

This takes us to the issue of having machines or software programs, which are capable to do that, i.e. translate for us. Machine Translation has been a topic of interest of scientists even before the invention of digital computers and was one of the first problems of computational linguistic investigated.

First attempts of Machine Translation were based on bilingual dictionaries and grammatical rules for reordering words in the target language (Hutchins, 2005). After the first attempts in 1950s it was thought that Machine Translation would be a solved problem in few years but it was soon realized that MT is much more than looking words in dictionary and ordering them according to some simple grammatical rules. There are even debates that Machine Translation is not even possible because translating a literal work is in many ways comparable with the original work, i.e. translation not only requires good knowledge of grammar and vocabulary of both target and source language but also requires a good grasp on the domain of the text being translated e.g. if a scientific paper is to be translated then a translator not having a good knowledge of that field might not be able to fully transfer the idea presented in the original text into the translation.

Moreover utterances in natural languages have inherent ambiguities i.e. certain words, phrases or even sentences would mean different things depending on the context e.g. idioms or simple assertions. Translation therefore requires deep understanding of cultural and communication norms of the speakers of the language i.e. how speakers of both target and source language use their language in different scenarios what particular style or construction they use to express certain concepts in certain situation.

It is due to these issues; the practical aim of Machine Translation is not to produce 100% human like translations, although it was the original idea and still the dream, but to produce an economical, fast and convenient translation of text for the required purposes. The translation produced by the computer can be used to get a vague idea of the message conveyed or it can also be used as primarily translation to be post edited by a person to produce a high quality equivalent text in target language.

Since MT is a heavily researched field therefore people working in the field has developed many techniques and have tried to approach the problem by many different angles.

1.2 Approaches to Machine Translation

Approaches to MT are generally categories into these categories.

- Rule-based.
- Example-Based.
- Statistical.
- Hybrid.

These techniques not only differ in the way the problem of MT is perceived but also the level or depth of linguistic analysis done.

1.2.1 Rule-Based

Dictionary-based machine translation: As the name suggests this approach translates as a dictionary does i.e. word by word. In this approach there is usually not much correlation of meaning. Dictionary look ups can also be done with lemmatization and morphological analysis. There can also be simple grammatical rules to reorder the sentence according to the target language but usually this most simple MT approach is used to get translation of words and phrases rather than complete sentences.

Transfer-base MT: Transfer-based MT is a category of rule-based MT which relies on analyzing the source sentence up to a certain level; the level could be shallow i.e. Syntactic analysis or it can be deep i.e. semantic analysis, and converting the source sentence into its formal representation. After converting into this formal representation the sentence is then transferred into its equivalent formal representation in target language according to some hand crafted or machine learned rules.

Interlingua-based machine translation: Interlingua-based machine translation is similar to Transfer-based MT in a way that it also analyzes the source sentence and transfer it into some intermediate formal representation, the analysis is deep i.e. analysis is done up to the semantic level. But the difference from Transfer-base MT is that the formal intermediate representation is language independent, moreover rules for converting into target language are not based on transferring from source formal representation to target formal representation, but rules are written to convert from the language independent intermediate representation to target language. One advantage is that there is no need to have transformation rules for each pair of languages but disadvantage is that, to a create formalism capable of representing meanings of sentences of a wide domain is very difficult or almost impossible. Therefore Interlingua-based approaches are only used for multi lingual translation systems in restricted domains, though they are not fully functional systems but are research oriented prototypes.

1.2.2 Example-Based

Example-Based MT is based on the idea that translation is not done by linguistically analyzing the sentence but it is done by dividing the sentence into number of phrases and then translating these phrases individually and then combining these small parts to get the target sentence.

Example-Based MT relies on a large bilingual sentence aligned corpora of source and target language to learn translation of different chunks or phrases. Then it uses those learnt translations to translate a given source sentence in to target language.

1.2.3 Statistical Machine Translation

Statistical Machine Translation is based on the idea of noisy channel model (Manning, et al., 1999) i.e. it is assumed that the target language sentence was converted in to source language model due to some noise. The task of translation is to rediscover the target sentence given the source sentence.

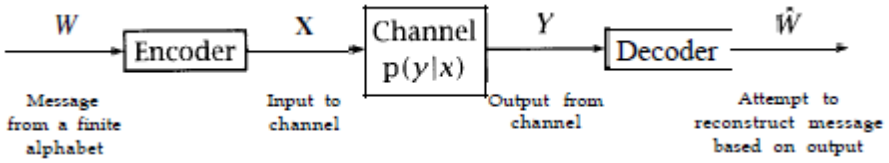


Figure 1 The Noisy Channel Model

The main equation of SMT is

$$\tilde{e} = arg\max_{e \in e^*} p(e|f) = arg\max_{e \in e^*} p(f|e)p(e)$$

Equation 1 SMT main Equation

e is the sentence in target language and f is the sentence in source language. The idea is to find the highest probable translation of e given f i.e. find the target sentence e from which f is most likely to come from i.e. after passing through the noisy channel. The SMT systems have typically three main modules. They are language model, translation model and a decoder.

Language model, which is the p(e) part of the equation. Its job is to provide how often a particular sentence appeared in the target language. Language model is trained by using a monolingual corpus of target language sentences.

Second main module i.e. the translation model or alignment model, which is the p(f | e) part of the equation. Its job is to show the probability that how often we see a phrase or word of f given that phrase or word of e is seen. This translation or alignment of phrases or words of f and e are learnt automatically by using bilingual parallel corpora of source and target languages utilizing methods developed by IBM called IBM model 1 to 6.

The final main module of an SMT system is the decoder. Its job is to find what is mentioned in the above equation i.e. the maximum probability that e can be a translation of f, or from the noisy channel perspective the highest probability, that it is e, which was converted to f, after going through the noisy channel. The problem of finding such an 'e' is an NP hard search problem since the search space is huge because of the size of the language and alignment models. It is for this reason the decoder used certain heuristic search algorithm maintaining a compromise between time and quality.

Given that a bilingual parallel corpus of source and target language is available

setting a SMT system is quick and economical since the necessary tools required to do so are not only well developed but also are freely available. The ongoing research in the field of SMT and availability of high performing hardware on an economical cost has made SMT systems the best performing systems so far i.e. SMT system outperforms systems based on other approaches, making SMT the most widely used and researched approach in MT.

1.2.4 Hybrid Systems

Since all approaches have their strengths and weaknesses, therefore it is natural to look for a way which can use both rule-based and statistical approaches. Since example base techniques can also be vaguely categories into automated rules base approaches depending upon how the translation rules learned are stored.

Hybrid systems are based on that very same idea that using both rules and statistics in translating from source to target language.

Hybrid systems have two broad natures i.e.

Rules engines output post-processed by statistics: in this approach the main task is done by rules engine and then statistics are used to adjust or even correct the output.

Statistical systems guided by rules. In this approach main job is done by statistical approach but some sort of post or pre processing e.g. normalization etc is done by rules. This approach is better than the previous hybrid approach since its offers more power and flexibility.

1.3 Combination of MT Systems Output

1.3.1 Problems with MT Systems

As mentioned above there are many approaches to Machine Translation. Although Statistical MT systems currently produce better results than others but they suffer from some problems, which are better dealt in different approaches. Same is true for other approaches i.e. while each approach has its prospects and consequences, strengths of one approach are the weaknesses of other and vice versa.

For instance rule-based systems are better in producing grammatically correct sentences (Thurmair, 2004) because rules are crafted by professional linguists, thereby making their output grammatically sound and more predictable. The advantage of this approach is that it is not domain dependent i.e. grammar of a language remain more or less same regardless of the domain in which the

utterances are spoken or written, therefore rule-based MT systems tends to produce better grammatically correct utterances no matter what domain the utterance belongs to.

Another strength of rule-based (Thurmair, 2004) MT systems is that they can work well even if the grammars of source and target language are very different from each other as rule-based MT works by analyzing the source sentence in to an intermediate form and then generating the sentence into the target language therefore it is possible to transfer from source structure to target structure regardless of the difference between the two, thanks to fact that both structures and their transfer rules are detailed defined.

However due to their dependence on carefully crafted detailed rules and reasonably large, specially designed sophisticated dictionaries, rule-based systems suffer from some problems. Rule-based systems are supposedly less robust i.e. even with large rules and dictionaries not all aspects of language can be modeled, therefore for many utterances, rule-based systems are unable to produce correct parses i.e. to analyze them correctly or at all, moreover for morphologically rich languages it is difficult for the rule-based system to make correct choices all the time.

Humans have tendency to produce sentences which are not grammatically perfect or even correct but are still understandable moreover in many situations or domains the general rules of the language are compromised and often used differently, many a times words, phrases or even sentences could mean totally different than their usual meaning depending on the context. These are the situations where rule-based systems suffer because of their dependence on rules and dictionaries i.e. either they can produce very good translations or they can produce something totally wrong.

While weakness of rule-based systems is the strength of Statistical MT systems. Since they are dependent on parallel corpus of source and target language, it provides them robustness i.e. in some situations they may not be able to produce perfect output but they will produce something which can be understood by the reader.

With a good quality and reasonable size corpus, Statistical MT systems produce quality translations. This is especially true for utterances of a specific domain i.e. terms, phrases and expressions specific to that domain can be better translated by statistical systems regardless of their grammatical quality provided that there are reasonable examples of them in the corpus.

But statistical systems suffer performance degradation if the utterances are out of domain i.e. if the corpus does not contain similar examples. Statistical systems can only work with things found in the corpus, since they depend on the frequency of occurrence in the corpus therefore expressions less frequent have less chance of being correctly translated even if they are linguistically

correct.

Moreover while statistical systems translate chunks of the source sentence to target sentence they tend to keep the constituent order of the source side resulting to be less grammatically correct if there are significant differences in the grammar of source and target language.

The above discussion leads us to the conclusion that no single approach is sufficient to produce high quality translation in all situations i.e. as both major approaches has their own strengths and weaknesses they both can prove to be more suitable in different situations while having some limitations.

It is natural to look for ways to overcome the limitations of MT systems.

1.3.2 System Output Combination as a Solution:

As it is concluded above no single approach or system works all the time. Therefore there is a need to find ways which can materialize strengths of different approaches while avoiding their weaknesses. Hybrid systems certainly are a solution and they are found to have good potential.

The fact is that with ongoing research the sharp distinction between rule-based and statistical system is fading away i.e. both commercial and research systems no matter statistical or rule-based utilizes many techniques belonging to the other approach to overcome certain limitations. Even if we characterize them as belonging to a certain category system are getting more hybrid in nature. Different systems perform better in some situations while others outperform them in other situations.

No matter a system is purely statistical, purely rule-based or is hybrid to any degree, it is a fact that no single system performs well all the time and different systems produce different translations of varying quality. Even systems based on same approach may produce different translation since they might be train on different data sets. It is often the case that some sentences are translated better by one system while other better by other systems. It is also possible that within a single sentence a chunk of it might be better translated by one system while some other chunk of it by some other system.

Nowadays one can have access to many diverse MT systems, all having their own strengths and weaknesses therefore it is practical to look for a way to get the best out from all of them i.e. to keep the best of their output while ignoring what is not good. Putting that another way; getting some parts of the output and combining them with parts of output of some other system to get an overall better output. This is the very definition of MT system combination.

MT system combination is an increasing trend (Burch, et al., 2010) in MT research community. The idea is to combine different MT systems to produce better translation than all individual systems thereby capitalizing on their

strengths and avoiding their limitations. And it has proved to have good potential. With the availability of many MT systems, system combination is getting a lot of attention in research (Burch, et al., 2010). There have been many approaches to system combination producing significant improvement over individual systems.

MT system combination has many promising applications. One can use system combination techniques to get a better translation from the outputs of many systems one have access to. As ongoing research in MT is so diverse and everyday people come up with new solutions to cope with certain problems of Machine Translation, and with the wide reach of the web, now it is possible to have access to many diverse MT systems at no or very reasonable cost. Despite the advancement in the field, there is no MT system which can be in informal terms, regarded as perfect or seem to work always, but all of them having their own areas of strength do something better than the other. This is certainly true for Statistical MT systems because of their dependence on the corpus they are trained on, they perform better in the area in which utterances of the corpus belongs to. With this scenario it is very beneficial from an application point of view to have techniques which can extract from all the translations produce by such diverse systems, a better one. But research in system combination can give us more benefits.

From research point of view combination techniques can give us valuable information about improving individual systems. As it will be explained in detail in the next chapter many combination techniques utilize lot of translation process information from participating systems. This information while helping in producing better results also helps in finding why a system fails to produce better results on certain situations and since information from system doing better in that particular situations is also there, so this information can be used to improve the design of that system. Such insight can help in designing a better more effective system which could or could not be somewhat hybrid in nature, which is certainly becoming a goal of Machine Translation research.

1.3.3 Our Task

The goal of this thesis is to investigate the System Combination in general and to review the techniques and research done in this field so far and also to experiment with certain techniques to come up with more effective methods for system combination. By effective it is meant that the resulting components based on those techniques will be as much language independent and individual systems independent as possible, i.e. those components will be able to work on different language pairs and can combine outputs from different combinations of participating systems having no constraints on the information they revealed about their process.

This thesis will not only discuss current techniques of the topic but it will also discuss the experimentation done on those techniques during this thesis. While most of the experiments done in this thesis are based on techniques already used in the field, but these experiments are more than just mere imitations of already done techniques because they are not only customized according the current scenario of the experimentation of the thesis but they also contain something of their own.

As it will be discussed in more detail in the literature review chapter that system combination can be just a re arrangement of the pieces of translation from participating systems without knowing much about how those translation were produced or it can be a lot of processing while utilizing a lot of information about the translation process of those system and practically redoing the translation process again to produce new translation.

Having insight information of translation process of a system is not always possible as in the case of commercially available systems. Even if it is available, since systems are diverse in nature, information provided by them might not be coherent with each other causing one not only to devise a strategy to extract coherent information out of them, but also causing the system do to some extra processing, and also making the technique dependent on particular system as opposed to any combination of systems. This extra effort can be avoided by providing detailed specification about the information which the participating system should provide thereby constraining the systems to produce certain output and also limiting the applicability of the technique. While the in-depth approaches might have some research oriented benefits they certainly lack application oriented potentials.

While techniques just relying on outputs of participating systems have many applicability advantages such as they can work with any combination of systems available which is quite an advantage as one can have access to the output of many systems.

It is for these reasons; this thesis concentrates upon techniques relying only on translation produce by systems without bothering much about how they were produced, because of the broad applicability potentials of these techniques in a wide range of scenario. Moreover since the other i.e. the in-depth approach can raise question about whether it is beneficial to implement the solution as individual system external to participating system or to it is better to change those individual systems to address those shortcomings thereby causing a divergence from this *topic* or even questioning the need of this topic altogether.

2 State of the Art in MT Combination

This chapter describes the state of the art of MT combination. I.e. it describes in detail the different types of techniques used so far by researchers for combining Machine Translation systems. MT system combination is receiving a lot of attention from the research community (Burch, et al., 2010)because of its potential benefits by utilizing many diverse MT systems.

2.1 Black box and White Box Approaches

MT combination can work, only with translation produced by participating systems i.e. combination technique does not have, or it require any insight information about the translation process of individual systems. All what the combining system does is that, it takes the outputs from the systems and uses it as its input and does some processing upon them. This is usually some sort of aligning and re scoring of the translations to produce new output translation which is supposedly better than the output of any of the individual system.

Such an approach is called black box approach. As the name suggests the combination system sees the individual systems as black boxes. Which produce some output and what goes inside the black boxes is unknown to the combination system and has no effect on the processing or decision making of the combination approach.

As opposed to black box approaches are in-depth approaches which in addition to the outputs from individual systems also use information regarding the translation process of those systems. These approaches are called Glass box approaches. Glass box approaches usually use information such as word translation pairs or phrase translation pairs and decoding lattices of participating translation systems. Glass box approaches while being significant from a research point of view because they give insight of the translation process, are limiting in number of ways. They require systems to give additional information regarding their translation process which is not

possible in many cases especially in the case of commercially available systems. Even in the case of availability, extracting coherent information from diverse systems is a resource consuming task.

Because of the above mentioned complexities, glass box techniques are hard to setup primarily because of the issue of availability of information required. Even with the availability of required information, glass box approaches requires additional processing for making the information usable by the combination technique or enforce participating systems to reveal certain information in a particular fashion. This limits their usage just for some particular systems and scenarios. It is for this reason glass box techniques are not used very widely for MT system combination research.

On the other hand black box approaches are the most common and widely used combination techniques and therefore are the primary focus of this thesis. Because of this reason this literature review chapter, for the most part will discuss black box techniques but glass box approaches will also be reviewed for the sake of completeness.

2.2 Some Glass box Approaches

This section briefly describes some glass box techniques for MT system combination. Glass box approaches utilize some additional information regarding the translation process of systems. That information normally includes different model scores from the systems, phrase and word probability, list of alternate translations of source words or phrases and some information about lattice creation by the individual system decoders.

Some glass box techniques include the one used by (Specia, et al., 2009) in which they use white box technique to extract translation process information from participating systems. Information they extracted is model scores, word phrase probability and alternate translation of source words. The authors then used a learning algorithm based on regression analysis to evaluate the sentence level quality of participating systems. They did not use that information for System combination directly. The reason why this approach is mentioned here is that this evaluation mechanism can be used in MT combination for re-ranking the hypothesis and selecting the best translation candidate.

A Glass box approach purely used for MT system combination is described by (Nirenburg, et al., 1994). In this approach the authors use three different types of MT systems namely one knowledge based MT system, one example base and a lexical transfer MT system. What the authors do is to take target phrases translation from each system together with some additional information about them such as their quality score etc. And they store all that in a chart

like data structure. Then they use a chart walk algorithm to select the best combination of edges from overall collection of candidate edges and reorder them to produce the final translation.

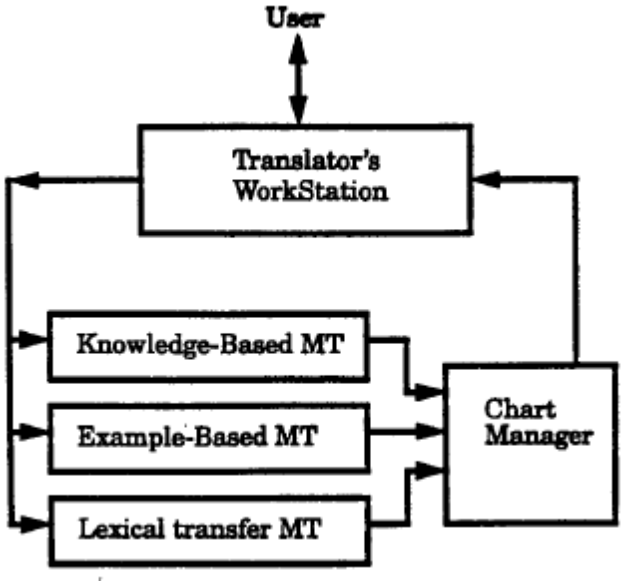


Figure 2: Multi Engine Translator

Another approach is (Huang, et al., 2007). What the authors do in this approach is to make each system provide, in addition to its 1-best translation output, detailed information about their process of translation such as what target word was generated for a certain source word and its order etc. Each participating system provides this information in an XML file.

This information is augmented into a phrase translation table, which the authors refer to as, training the combination system with test specific data. The table is then pruned and the source sentence is decoded using this pruned phrase translation table, by the selected decoder to produce a new translation, which in turn is better than all individual systems translations.

2.3 Black Box Approaches

Black box approaches due to their simple requirement of just using the output from individual systems are the most widely used MT combination techniques for research. Over the years, due to the development of many diverse MT systems build on diverse paradigms and easy availability of these systems resulted in lot of focus on system combination, bulk of which is on black box approaches.

Over the years a large number of black box combination techniques have been tried by researchers. Combination of output on all levels has been tried. Although bulk of the research effort has been on different types of Confusion

Network decoding, but researchers has also experimented with other techniques i.e. combination of outputs on levels, higher than that of words such as sentence and phrase level combination.

Though it will not be exaggerating to say that, the current research in system combination is all about Confusion Network or lattice decoding since a vast majority of research is about them, and issue related to the problems faced by Confusion Network decoding and their optimization. But other techniques of combination will also be discussed in this literature review.

2.3.1 Sentence Level Combination

Sentence level combination techniques are one of the simplest techniques used for MT system combination. Since these techniques work at sentence level, the prospects of having an improvement in translation quality are quite limited.

In general sentence level combination techniques are all about taking outputs from participating systems and re-ranking them on some criterion. Those criterions are usually scoring by a language model in addition to some other features related to those sentences. Different techniques usually differ from each other in terms of choice of features and how those used features and language model are combined to score the candidate sentence. The most common is using a liner combination of them. The idea is to pick the best sentence among the sentences produced by individual systems. As is the cases that some sentences are translated better by some systems while other by other systems, so choosing the best one, among all candidate sentence translations, gives an overall increase in translation quality of the test set.

While the margin for improvement at this level is quite small since the best chosen option is expected to most probably come from the best performing system among the participating ones. Therefore sentence level techniques have limited usage. The most interesting use of sentence level techniques done by researchers is to use these techniques in addition to more sophisticated MT combination techniques i.e. to first use a more sophisticated combination technique to produce n outputs and then re-ranking them with sentence level combination techniques.

(Rosti, et al., 2007) use a sentence level combination approach which uses N best list from all participating systems. Usually only one best output is available from systems but Sentence level techniques work better with more candidate translations. The authors re-rank the merged N-best list from the scores from log-linear combination of scores from, a 5gram Language Model, number of words in the candidate translation and confidence score. The confidence is calculated by features such as the number of systems generating that hypothesis and sentence posterior of that hypothesis etc. the authors

report that on its own this technique did not gave them much improvement but when used in addition to other techniques it proved to be useful.

Same is the case with the sentence level technique used by (Huang, et al., 2007). The authors use sentence level combination by re-ranking hypothesis list by two feature functions. For the first function they used a 5-gram language model to re-score hypothesis, the other feature function they tried was based on a 5gram language model which was calculated on a mixed stream of words and Parts of speech tags. They also used this technique in addition to the other more sophisticated techniques they used for their experiments.

(Specia, et al., 2009) While concentrating on glass box technique to evaluate the sentences, they also experimented with feature selected with black box techniques to evaluate candidate sentences.

2.3.2 Phrase Level Combination

As the name suggests Phrase level combination techniques utilize combination at phrase level rather than sentence level. This gives these techniques more margins for improvement since there are more options to choose from rather than just the list of candidate translations.

In general, phrase level combination techniques are about extracting new phrase tables from the list of candidate sentence translations and then using that newly created phrase table and a selected decoder to decode a new translation of the source sentence.

Phrase level combination techniques can both be black box and glass box in nature. As mentioned in section2.2 that the participating systems can provide source and target alignments used for the translation of the sentence. In case if such information is not available for the participating systems as is often the case then these alignments can be learned automatically by alignments tool such as Giza++ thus making phrase level combination techniques also, black boxed in nature.

Assuming that the new phrase tables are generated by a tool or make available by any other mean, source sentence are re-translated in manner described in (Huang, et al., 2007) and (Rosti, et al., 2007) etc.

Phrase level techniques like sentence level techniques also have limited usage and the most commonly used and the most widely researched techniques are word level combination techniques or better known as Confusion Network decoding techniques described in next section.

2.3.3 Word level or Confusion Network Decoding Techniques

Word level combination or better known as Confusion Network decoding

techniques are the most widely used and researched combination techniques. Almost all the research going on in Combination of MT field is upon these approaches.

In general Confusion Network decoding based techniques consists of creating a Confusion Network from the words of output sentences generated by participating systems. After creating that Confusion Network, the best path is searched in it according to some scoring technique to produce a new translation of the sentence, which is supposedly better than any of the participating systems.

There are number of issues which need to be addressed. First one is to find a way for effective creation of the CN. Translation systems produce outputs in different word orders therefore it is a quite challenging task to align those outputs to see on which places, words from different system outputs, are reinforcing each other and on which positions, they are providing alternates. A lot of research has been done in this regard i.e. to find effective alignments techniques to create a better Confusion Network.

Over the years researchers have come up with the number of different techniques to align system output to build a more effective Confusion Network which can be efficiently searched to produce an improved translation.

These techniques can be broadly classified as skeleton based and non skeleton based techniques. Variation is not only present in ways words are aligned to create CNs but researchers also experiment with different searching techniques to get a better translation out of the created CN.

Skeleton Based Techniques. In these types of techniques CN is built around a skeleton or backbone sentence. One of the outputs from participating systems is selected as skeleton or backbone of the CN, while all other system outputs are aligned to the backbone to create the CN. The reason why a backbone or skeleton needs to be selected is because MT systems can generate outputs in any word order while selecting the skeleton helps in determining the word order for the translation. That is why, careful selection of the skeleton is one the crucial problem in CN decoding techniques.

There are various ways in which the skeleton can be selected. The simplest being to select the best performing system's output to be the skeleton but there are also more sophisticated techniques used by researchers. (Sim, et al., 2007) use Minimum Bayes Risk (MBR) decoding to select the skeleton instead of choosing the best system output. What the authors do is that they measure the MBR as expected loss over the posterior probability distribution. The problem with this approach is its cost of computing increases quadratically with the increase in the size of N-best list. (Rosti, et al., 2007) Select the hypothesis as skeleton which best agrees with other hypothesis on an average.

It is an approach somewhat similar to MBR approach used by (Sim, et al., 2007).

After choosing a skeleton, different alignment techniques are used to build the CN, for example (Bangalore, et al., 2001) use WER edit distance alignments to build a monotone CN around a selected skeleton. This approach while giving some prospects is quite limited because of its monotone nature. Because it relies too much upon the word ordering of the skeleton therefore the room for improvement is limited. (Sim, et al., 2007) use Translation Edit Rate (TER) alignments to build a CN. What TER does is, it measure the minimum number of edits between sentences. By this edit measurement the alignment between sentences is also determined. This approach gives more flexibility in word ordering than the WER approach. (Karakos, et al., 2008) Experimented with Inversion Transduction Grammar ITG formalism to aligned outputs which overcome certain drawbacks of the TER approach and also creates CN in a more efficient manner.

(Matusov, et al., 2006) used a different approach to get the alignment. They aligned the outputs from the participating systems, similar to the fashion done in the training process of a statistical MT system. The only difference is that aligned sentences are in the same language rather than being different as is the case in SMT. The authors use GIZA++ alignment tool for this purpose. This approach gives them more flexibility as compared to previously described alignments techniques, but the drawback of this approach is that it requires a lot of training upon a development set.

Well building the CN is just one of the tasks; next challenge is to decode it to get an output from it. A lot of research has been done in finding ways to effectively find the best path and there are lots of variations tried by researchers to do that. Generally the best path is chosen by scoring the candidate paths on language models and other voting schemes and feature functions.

After the creation of the CN each arc is given the posterior probability of its label i.e. the word which is there. That probability is proportional to the number of systems producing that word. Putting another way, if the same word is aligned at a certain position within the CN, then it has a higher probability of appearing in the final output. Output is then produced by finding the path which leads to maximum probability.

(Rosti, et al., 2007) use this probability assignment procedure and then use a lattice decoding algorithm to generate N best list from the CN and then re-rank it by scoring the list by some feature functions to get the final output. Once the CN is created all techniques use some kind of standard lattice decoding algorithms to get the final path.

(He, et al., 2009) use a joint optimization technique to eliminate the need for

choosing a skeleton. What the authors do is, they create and decode the CN jointly using a log linear model instead of performing these tasks separately.

Multiple or no Skeleton based Techniques. As it is discussed that word ordering plays a crucial role in MT quality therefore some sort of word ordering has to be decided for the combination techniques. But it has been found out that no matter how good the selection of the skeleton of CN is there are still limitations to these approaches. Even the selection of skeleton is a big problem on its own. If we assumed that the best output is selected to be the skeleton even then the margin of improvement is limited. Because the CN is too much dependent upon the skeletons word order and it is possible that the skeleton may not have the suitable word order for that particular translation.

In order to overcome this limitation, number of techniques has been used e.g. (Heafield, et al., 2009) instead of relying on a single output to be the skeleton, change the effective skeleton on a phrase to phrase basis. Not sticking to a particular output sentence gives there method more freedom in terms of word orders.

Although having a phrase based skeleton gives some freedom but still it remains to be a single CN with limited paths to choose from i.e. the word order is still somewhat limited. Other approaches to enhance the option of path selection are, to use each system output as backbone in turn and thereby creating many CNs. (Heafield, et al., 2010) use this approach to create many CNs by taking every hypothesis as back bone each time. After creating the CNs, their algorithm starts searching the path from the beginning of any of the CNs and it can switch between CNs at any time, but taking into account not to duplicate words during switching. This switching between CNs techniques can be taken as to be a lattice decoding technique. (Matusov, et al., 2006) use a similar idea of using more than one CNs to create a lattice and use a voting scheme based on union of the CNs to get the final output path.

No matter how the skeleton is chosen and no matter if we are using one or many CNs. Confusion Networks remain limited in a sense that the nodes between CN has 1 to 1 mapping. Confusion Networks limits that a word is aligned to another word but in languages it is possible that a group of words or a phrase is aligned to another phrase or just a word. Though CNs try to solve this problem by aligning words to empty words but it creates chances to generate wrong or missing paths. Therefore there is a need to have a technique which allows arbitrary mappings between hypothesis alignments. (Feng, et al., 2009) use a lattice instead of a CN which provides this ability. What the authors do is to get alignments from the hypothesis sentence using indirect-HMM based methods and extract phrase pairs from them after normalizing those alignments. They then create a lattice in the light of those extracted phrase pairs. The generated lattice has arbitrary mappings and it is

decoded using standard log-linear techniques to produce the final output.

As it is clear from the above review of MT combination literature, there are numbers of diverse techniques for MT combination. Which one is the best; there is not a clear cut answer to this question. All of the above mentioned techniques have their prospects and consequences. Each of the techniques described above has its own application areas i.e. each technique will perform better than other in certain scenarios. Therefore selection of using a particular approach depends upon many factors such as what will be the applicability of the combination application and what input will be available for the application to work upon and under what environment the application will perform whether it will have some time constraints or not etc.

As it is mentioned in the task statement, the purpose of this thesis is to come up with techniques which are applicable on a general scale therefore all experimentation of this thesis is based on techniques which have more applied potential.

3 Environment and Tools used for Experimentation

This chapter briefly describes different tools utilized for the experiments done for this thesis.

3.1 Operating System and Programming Language

The operating system on which all the tools installed and components developed for this thesis is Ubuntu distribution of Linux. Being an open source OS Linux is the choice of all researchers and most of the components which are required for this work are all developed on Linux platform. Though almost all of them can be run upon other Operating systems such as windows but with its powerful bash scripts, Linux based OS are ideal environments for linguistic related development.

All the programming for this thesis is done using Perl as programming language. Perl is completely open source language and it has an easy to learn syntax. Perl was designed with a mind set to make its programs easier to understand by humans rather easier to understand by computers. Because of this and the fact that it is very portable and has a large number of features make it a widely used language especially for research purposes.

Perl's easy to use syntax, its data structures, built in functions and powerful regular expression processing makes it a very productive language for linguistic related task. It is one of the widely used languages for natural language processing tasks and for this reason is the programming language for this thesis.

3.2 SRI Language Modeling Toolkit

SRILM is a set of tools used for statistical language modeling. The statistical LMs built by this toolkit are used for speech recognition, tagging and segmentation and machine translation.

Among the large list of tools available in this toolkit, the ones used for this

thesis are.

- ngram-count
- ngram
- lattice-tool

3.2.1 Ngram-count

This program creates n-gram language models and stores them in ARPA n-gram or in binary format. It is capable of creating a variety of LM with all state of the art smoothing and discounting techniques. All the language models used in this thesis are built using this tool.

3.2.2 Ngram

This tool can perform a variety of tasks on n-gram language models. Its most common tasks are sentence scoring, perplexity calculation and sentence generation. It can also work with more than one language models and perform various operations with them such as model interpolation etc. another interesting feature of this tool is that it can work as a server listening to request at a port.

All language model scores in the experiments of this thesis are computed using this tool.

3.2.3 Lattice-tool

This tool is capable of performing various tasks on a lattice or a Confusion Network since a CN is also a particular type of Lattice. The main operations it can perform are size reduction of the lattice, pruning, weight assignment and the most useful for our case, decoding of the best hypothesis.

This tool has been used in the experiments for this thesis to find the best path in the Confusion Network using viterbi search and n-gram language model.

3.3 METEOR

METEOR is a tool which is used for performing many NLP related tasks. Its main task is to evaluate machine translation hypothesis against a reference translation based on a similarity scored calculated based upon the alignments between those sentences. Although it is an evaluation metric but it is also used as an alignment tool since it does evaluation by alignment.

METEOR is used in the experimentation of this thesis as an alignment tool to align different hypothesis from participating systems in order to build the

Confusion Network.

3.4 HTK

HTK toolkit is set of tools for creating HMM models for Speech Recognition systems. The reason why it is included in the thesis on MT system combination is that it has a series of tools that can be used in other areas.

Hparse

This tool is used to create lattice files according to DARPA format, the format which lattice-tool of SRILM toolkit works on. This tool has been used in the experiments to convert the built Confusion Network into lattice-tool compatible format.

4 Language Modeling & Others Details about Experiments

This chapter describes in detail the data used for the experimentation of this thesis. It also describes the pair of languages used for testing and characteristics of those languages. Descriptions of language models used in the work, and individual systems used for combination are also provided.

4.1 Translation Direction

Though the methods used in this thesis are all potentially useful for any language pairs, and they can work for any combination of MT systems, since all they require is the system output. But for testing purposes Translation from Czech to English is used. Below is a brief description of both languages.

4.1.1 Czech Language

Czech language which historically has also been called Bohemian is the member of West Slavic family which is the subfamily of Indo-European family of languages. It has approximately 12 million native speakers. Czech language has seven cases, has a rich morphology and because it requires a lot of agreement between verbs and other constituents of the sentence, it is a free order and pro-drop language.

There has been a lot of NLP research done on Czech language. Czech also attracts a lot of attention in Machine Translation research and is included as a language for evaluation in many major MT evaluation campaigns and projects like ACL joint workshop on machine translation (Burch, et al., 2010), Euro Matrix project and other major European Projects.

4.1.2 English Language

English belongs to the family of West Germanic languages. It was developed in England during the Anglo-Saxon era. As a result of the military, economic, scientific, political, and cultural influence of the British colonialism during the 18th, 19th, and early 20th centuries, and of the United States since the middle

of last century, it has become a major language or in other terms the *lingua franca* in many parts of the world.

Because it is globally used language therefore English is one of the most researched languages in the field of NLP and Machine Translation.

The reason why English is used as a Target language for the experimentation of this thesis is to have an insight look at the translation produced and to easily analyze them and not to have full reliance on automatic evaluation metrics such as BLEU etc and also for the fact that they are lot of linguistic resources easily available for Language Modeling etc.

4.2 Data Used for LM Creation

This section describes the corpora used for the language models utilized in the experiments of this thesis.

As English is used as the target language and all techniques used for experimentation are so called Black boxed in nature therefore only language models of English were required and created.

English Portions of CzEng, UN and EuroParl corpus is utilized. Below is a brief description of them.

4.2.1 CzEng Corpus

CzEng (Bojar, et al., 2009)corpus is a sentence-parallel corpus of Czech and English language and is currently in its 0.9 version. It is compiled at the Institute of Formal and Applied Linguistics (ÚFAL). UFAL is part of the Faculty of Mathematics and Physics of Charles University Prague.

Current version of CzEng contains approximately 8 million parallel sentences and has approximately 93 million English and 82 million Czech tokens. CzEng covers variety of genres like news, movies, legal, technical and web etc.

The reason why CzEng covers so many domains and genres is due to the diverse sources it get its data from. One of the major sources of this corpus is movie subtitles from different internet subtitle archives which contribute approximately 3.5 million sentences. CzEng also uses many factious eBooks and they contribute approximately 1 million sentences. Others sources include European Union legal documents, Czech news portals, some localization documents of brands like gnome, Microsoft etc, Kačenka corpus which is Czech-English corpus compiled by Masaryk University and manual translations of Wall street journal etc.

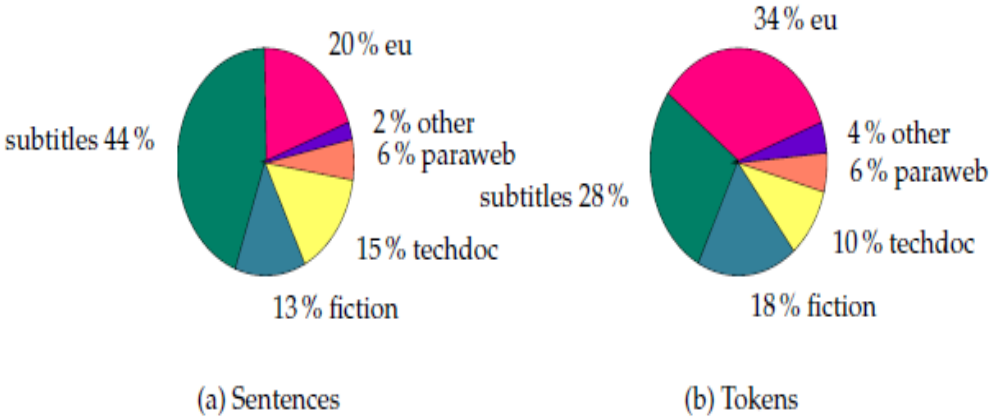


Figure 3 Percentage of different Generes in CZEng

CzEng corpus is more than just a parallel text corpus. The whole corpus is augmented by information regarding the analysis of both Czech and English sentences on morphological, analytical and tectogrammatical layers.

Needless to say, only the plain text of English side of the corpus is utilized for the experimentation of the thesis though the annotation information in the corpus can be valuable resource for glass box techniques.

4.2.2 Europarl Corpus

Europarl corpus is a multilingual parallel corpus of 11 European languages namely French, Italian, Spanish, Portuguese, English, Dutch, German, Danish, Swedish, Greek and Finnish. Being an Organization of culturally diverse countries with many different languages, the European Parliament proceedings are translated into many different languages. These proceedings are published on the European Parliaments website (Koehn, 2005). The website publishes these proceedings in the form of HTML files with speaker and other reference information.

Europarl corpus currently in its 5th version is compiled by crawling these proceedings and sentences are aligned by tools based on the Church and Gale algorithm. Since the proceedings cover a lot of topics therefore text is organized into chapters. It is released as parallel corpora of language pairs which include English i.e. Danish-English, Italian-English etc. It is a major resource for MT research for European languages.

English portion of Danish-English release was extracted for the language models used for this thesis. It contains approximately 1.6 million sentences and 46 million English tokens in the form of XML files which gave information such as chapter id, speaker id and paragraph etc.

4.2.3 UN French-English Corpus

United Nations much like European parliament is a global organization with diverse cultures and languages. It also translates its proceedings in many different languages. Frech-English corpus is based on these United Nations proceedings and is made available by LDC i.e. Language Data Consortium, which is a major contributor for MT related research.

The version of the corpus used in this thesis, consists of two plain text files, one containing English and the other containing French. Both files contain parallel sentences. The English file, the one of our interest contains approximately 7.2 million sentences and approximately 180 million tokens.

4.3 Language Models

For experimentation three language models are built. One is made from joining the text of CzEng and UN corpus and other from the UN and Europarl corpus. The first one is made from almost 15 million sentences while second was made of approximately 8.5 million sentences and third one was made just from CzEng alone using 6.4 million sentences.

Ngram-count tool of SRILM toolkit was used for this purpose. The order of all LMs is 3 i.e. all of them are trigram base LMs. Although higher order LM are supposedly better, but the added complexity is usually higher than the benefits they offer so it is often found that trigram LMs are the best compromise in performance and accuracy (Filippova, et al., 2009). As it is the goal of this thesis to find combination techniques which not only improve translation output but also are efficient in nature to be applicable in variety of application therefore trigram LM were chosen.

Data sparsity is always an issue in statistical LMs therefore smoothing always helps to perform better. It has been noted in the literature many times that Chen and Goodman's described method for Kneser-Ney discounting method with interpolation works best for smoothing and is widely used. It is why LMs used in this work are also smoothed using this setting.

4.4 Systems Used for Combination

The individual systems output used for the experimentation came from ACL joint workshop on Machine Translation (Burch, et al., 2010) which is one of the major MT evaluation campaigns and it major goal is to evaluate the state of the art in the field of MT. In the workshop there is a shared translation task and a combination task.

For the Translation task participant were required to translate a test set

comprising of news stories selected from variety of news sources such as iDNES.cz and BBC etc, and were professionally translated directly in to all languages of the task.

The primary submissions of the systems participating in the evaluation were made available for the system combination task, and are used for the experimentation of this thesis.

Systems which provided translation in Czech-English direction along with their BLEU scores are listed below.

All of the systems mentioned below provided their 1 best results in SGML format. No information regarding translation process is given but since experiments of the thesis are based on black box techniques therefore such information is not required.

No.	System Name	Developer	BLEU
1	AALTO	Aalto University, Finland	17.37
2	Cmu-cunei	Carnegie Mellon University	22.38
3	Cu-bojar	Dr Bojar of Charles University	19.18
4	Cu-zeman	Dr Zeman of Charles University	13.85
5	Google	Google Inc.	22.84
6	UEDIN	University of Edinburgh	22.28

Table 1: Individual Systems and their BLEU score.

5 Sentence Based Combination Experiments

5.1 Introduction

Sentenced based combination or in other terms sentence re-ranking techniques are based on choosing the best sentence among the outputs of participating systems, based on different scoring criterion. Sentences are re-ranked according to their quality, which is usually measured by scoring them against an n-gram language model, and by some other features.

5.2 Motivation

Being the simplest techniques of combination, these techniques cannot do much change in the output translation. In fact they do not make any changes to the output of individual systems giving them very limited opportunities to make positive changes. All they can do is to choose the best translation among the options.

The motive behind re-ranking sentences is, that translation is usually done for block of sentences e.g. news articles, web pages, manuals and other form of documents. It is a possibility that a system A translates some sentences in the document with very good quality but system A does not translate other's sentences with the same quality for any reason what so ever. E.g. sentences to be translated are very different from ones which were used for training the system etc. System B however translates those sentences better which were not so well translated by system A, and not do so well on those better dealt by system A or C, D etc. By choosing the best work of each system for all the sentences to be translated the overall translation quality of the document or test set can be improved.

Re-ranking techniques have many practical applications. One being mentioned in the above paragraph, i.e. improving the overall quality of sentences; other application is simply finding the best one among all candidates for any reason what so ever. Because of their simplicity they can be used in addition for other

combination techniques i.e. to re-rank the n-best list produced by other combination techniques (Rosti, et al., 2007) and (Huang, et al., 2007).

Re-ranking techniques are not limited to be used just for system combination. They can also be used with in a machine translation system utilizing any paradigm to simply rank its produced n-best list. All MT systems of course rank their produced n-best list but this ranking is only based on criterion used within the translation process, re-ranking with criterion other than the one in translation process can provide an extra edge in translation quality.

5.3 Scoring against just Trigram LM

In this series of experiments 1-best output of each system was used. Translation output from each system was scored and ranked against trigram LMs. For each sentence, translation output from each system was scored and the highest scoring hypothesis was chosen as the translation. For this series, no other features were used; scoring is done solely against language models.

5.3.1 Summing up Trigrams of Sentence

As the name suggests, in these experiments the candidate sentences were divided into trigrams and those individual trigrams were scored against a language model. The sum of the log probability of all the trigrams of a sentence was taken to get the sentence score.

Step by step Procedure:

1. Get the produced translation sets of all individual system.
2. For each sentence of the set, repeat step 3 to 7 to get the final output set.
3. Get the candidate translation of each system.
4. For each candidate translation divide it into trigrams.
5. Score trigrams against the language model using ngram tool.
6. Sum up the log probability of trigrams to get the sentence score.
7. Select the sentence having the highest score and put it into output set.

Results of the experiments are shown in the following table.

No.	System	BLUE
1	AALTO	17.37
2	Cmu-cunei	22.38
3	Cu-bojar	19.18
4	Cu-zeman	13.85
5	Google	22.84
6	UEDIN	22.28
7	comb_UNEU	18.31
8	comb_UNCzEng	17.40
9	comb_CzEng	17.31

Table 2: Score after scoring hypothesis with LM score only and summation of trigrams scores

Rows 1 to 6 are scores of the individual systems. While comb_UNEU is the experiment result by re-ranking the hypothesis list against the LM made from United Nations and europarl's corpus, comb_UNCzEng is from the LM made from joining UN and CzEng corpus and comb_CzEng is from the LM made solely out of CzEng corpus.

Results from all LM are lower than the best performing system. It can also be seen that among all LM UNEU got the best results for this particular test set. UNCzEng even though it is bigger but was less effective for this particular set maybe because being bigger in size also increased its sparsity. CzEng alone proved to be little less for this test set.

5.3.2 Summing up Trigrams of Sentence Without Marking Start and End

Ngram tool puts a special symbol at start and end of each string given to it as input. As in our case the trigrams are part of a sentence rather than sentences on their own. Therefore in this series of experiments, scoring was taken by instructing the ngram tool not to put start and end around the trigrams. Summation of log probability of trigrams scores was done the same way as is in the previous set.

Step by step Procedure:

1. Get the produced translation sets of all individual system.
2. Instruct ngram tool not to put start and end symbol around trigrams.
3. For each sentence of the set, repeat step 4 to 8 to get the final output set.
4. Get the candidate translation of each system.
5. For each candidate translation divide it into trigrams.

6. Score trigrams against the language model using ngram tool.
7. Sum up the log probability of trigrams to get the sentence score.
8. Select the sentence having the highest score and put it into output set.

Results of the experiments are shown in the following table.

No.	System	BLEU
1	AALTO	17.37
2	Cmu-cunei	22.38
3	Cu-bojar	19.18
4	Cu-zeman	13.85
5	Google	22.84
6	UEDIN	22.28
7	comb_UNEU	18.37
8	comb_UNCzEng	17.32
9	comb_CzEng	17.12

Table 3: BLEU score after scoring hypothesis with LM score only, and summation of trigrams scores without start end symbol insertion

Again rows 1 to 6 are scores of the individual systems with comb_UNEU being the experiment by re-ranking the hypothesis list against the LM made from United Nations and europarl's corpus. Experiment comb_UNCzEng is from the LM from UN and CzEng corpus and comb_CzEng only from the CzEng corpus.

Again all of the re-ranking schemes scored lower than the best performing system in term of BLEU. There is slight improvement in UNEU i.e. 18.37 instead of 18.31 of previous set but other LM score seems to drop, than their counterparts in the previous set e.g. comb_UNCzEng 17.32 compared to 17.40 in the previous set etc.

Although highest BLEU got in this set is a bit better than the previous one, but still not enough to get improvement over individual systems.

5.3.3 Scoring Whole Sentence Against LM

Summing up the scores of trigrams of a sentence is not proving to be a very effective technique. Ngram tool also gives the option of scoring a complete sentence. Instead of summing up the trigrams score, scoring technique used by ngram tool was utilized for this set of experiments. For each sentence the output sentence produced by each system was scored against LM and re-ranking and selecting the highest scoring hypothesis, as the translation, was done the same way as in previous sets.

Step by step Procedure:

1. Get the produced translation sets of all individual system.
2. For each sentence of the set, repeat step 3 to 5 to get the final output set.
3. Get the candidate translation of each system.
4. Score each candidate translation against the language model using ngram tool.
5. Select the candidate having the highest score and put it into the output set.

The table below gives the results for this set of experiments and the style is of presentation is the same as in previous two sets.

No.	System	BLEU
1	AALTO	17.37
2	Cmu-cunei	22.38
3	Cu-bojar	19.18
4	Cu-zeman	13.85
5	Google	22.84
6	UEDIN	22.28
7	comb_UNEU	18.91
8	comb_UNCzEng	18.41
9	comb_CzEng	18.02

Table 4:BLEU score by scoring whole sentence just by LM

Results show that using ngram's default scoring works better than summing up log probabilities of trigrams as scores are better than the previous sets, but still they fall shorter to the BLEU scores of the individual systems.

5.3.4 Analysis of LM scoring

Re-ranking of sentence hypothesis based solely on LM scoring did not give improvement over individual systems in terms of BLEU. Upon investigating the produced output, nearly all scoring schemes showed that sentences having shorter length were given more score by the language models. Sentences produced by systems having better BLEU score were longer in length, therefore were having a disadvantage in getting selected as the final translation. Therefore overall BLEU score of the produced set was lower than the best performing individual system.

It is clear that re-ranking solely based upon LM score is not enough to get improvement over individual system, there must be some additional criterion upon which sentence hypothesis should be scored.

5.4 Scoring against Liner Combination of Features

As language model alone proved to be insufficient for re-ranking therefore In this series of experiments sentence hypothesis were scored against a linear combination of language model score and some other features.

5.4.1 Experimentation with the First set of Features

For this series of experiments, linear combination of following features was used.

- Language model score
- Number of words in the candidate translation
- Number of systems producing that particular candidate translation.

Language Model Score:

LM score was calculated in the same way as described in section 5.3. All variations such as sum of the log probability of individual trigrams with or without special symbols and scoring the whole sentence at once were tried.

Number of Words:

The motivation behind putting the number of words as a feature is to address the problem faced in experiments of section 5.3. The idea is to compensate hypothesis having longer length, for the disadvantage they have in language model scoring thereby giving them more chances to get selected as final output.

Number of Systems:

The motivation behind putting number of systems producing the hypothesis as a feature is to give higher priority to a hypothesis if more than one system is producing it. Although this feature works better with systems giving their N-best translation list instead of 1-best or if the number of participating systems is larger but none the less this feature does not give any negative impact on hypothesis selection.

Linear combination of these features can be best described by the following equation.

$$score = \lambda_1 LMscore + \lambda_2 NOW + \lambda_3 \mathrm{NOS}$$

Equation 2 : Linear combination of features.

Where $\lambda 1$ to $\lambda 3$ are the weights assigned to each of the features. LMscore is

language model score, NOW is number of words in the hypothesis and NOS is number of systems producing that hypothesis. All feature scores are scaled and normalized in order to make them more consistent with each other.

Step by step Procedure:

1. Get the produced translation sets of all individual system.
2. For each sentence of the set, repeat step 3 to 8 to get the final output set.
3. Get the candidate translation of each system.
4. For each candidate translation calculate the LM score by using any of the methods described in section 5.3
5. For each candidate translation calculate its number of words and number of systems producing it.
6. Scale and normalize all the values to make them more consistent.
7. Calculate the final score of the candidate sentence according to Equation 2.
8. Select the candidate sentence having the highest score and put it into output set.

No.	System	BLEU
1	AALTO	17.37
2	Cmu-cunei	22.38
3	Cu-bojar	19.18
4	Cu-zeman	13.85
5	Google	22.84
6	UEDIN	22.28
7	comb_CzEng	18.05
8	comb_CzEng_sum	17.33
9	comb_CzEng_sum_no_se	16.94
10	comb_UNCzEng	18.58
11	comb_UNCzEng_sum	17.40
12	comb_UNCzEng_sum_no_se	17.39
13	comb_UNEU	19.23
14	comb_UNEU_sum	18.31
15	comb_UNEU_sum_no_se	18.21

Table 5: Experiment results of linear combination of the first set of features

First 6 rows are BLEU scores of individual systems provided again here to easily get the comparison. Comb indicates the experiment results by re-ranking the hypothesis. UNEU, UNCzEng etc describes the language models used for scoring, difference is that now language model score is not the sole criteria as other features are also taking part in scoring the hypothesis.

The absence of sum means that the LM score was taken by Ngram's default scoring i.e. same as the one described in section 5.3.3. Presence of sum indicates that sum of individual trigrams score was taken as LM score, same way as in section 5.3.1 . Similarly the presence of no_se means that ngram tool was directed not to put start or end symbols around trigrams as in section 5.3.2.

E.g. comb_UNEU means that ranking of hypothesis was done by linear combination of features and language model score was taken from the LM made from UN and europarl corpus using ngram tools default scoring technique. Similarly comb_CZEng_sum means language model made from CzEng corpus was used for LM scoring while the sum of trigram score were taken to measure the LM score.

There are lots of similarities with the previous series i.e. UNEU again performs better than other LM etc. The highest score in this series is better than previous one i.e. 19.23 compared to 18.91 of the previous series indicating that introduction of new features resulted in better selection of hypothesis. But still it falls shorter than the best performing system.

5.4.2 Experimentation with the Second set of Features

Though features described in the previous section give little improvement but they are not good enough. There is a need to find some better features for scoring hypothesis.

The feature NOS in Equation 2 while being useful is too strict and limited. As we commonly have only 1-best list of participating systems therefore the value of NOS will not have much variation unless the number of participating systems, is quite large. Moreover even if the output of two systems differs only by a single word, NOS will give the value 1 losing any advantage which can be achieved by other common words.

Common trigrams:

For the above reason a less strict feature is used in this series of experiments for sentence scoring. That feature is common trigrams in candidate translations.

For each trigram of a candidate translation, this feature looks for similar trigrams in other candidate translations. With more systems producing the same trigram, more score is given to that candidate translation. Score of a

candidate also increases with an increase in the number of such trigrams. I.e. a candidate having 5 trigrams produced by 5 different systems each will have a higher score than a candidate having 5 trigrams produced by 4 different systems each.

LM score and NOW

The other two features used in this series are same as in section 5.4.1 i.e. LM score and number of words in the candidate translation. Only difference is that as all previous experiments showed that UNEU always performs better than other LM therefore only this LM was used in these and all further experiments.

With the introduction of common trigrams feature, Equation 2 is changed to the following equation.

$$score = \lambda_1 LMscore + \lambda_2 NOW + \lambda_3 \mathrm{CT}$$

Equation 3: Linear combination of the second set of features

Where λ1to3, LMscore and NOW are same as in Equation 2, CT is common trigrams scores of the candidate.

In addition to Equation 3 following variations were also tried.

$$score = \lambda_1 LMscore + \lambda_2 \mathrm{CT}$$

Equation 4 : just LM score and common trigrams

And

$$score = \mathrm{CT}$$

Equation 5: just common trigrams score

Step by step Procedure:

1. Get the produced translation sets of all individual system.
2. For each sentence of the set, repeat following steps to get the final output set, skipping step 5 if using Equation 4 and step 4 and 5 if using Equation 5.
3. Get the candidate translation of each system.
4. For each candidate translation calculate the LM score by the method described in section 5.3
5. For each candidate translation calculate its number of words.
6. Divide the candidate sentence into trigrams and calculate the CT score.
7. Scale and normalize all the values to make them more consistent.
8. Calculate the final score of the candidate sentence using appropriate equation.

9. Select the candidate having the highest score and put it into output set.

Results of this series of experiments are presented in table below.

No.	System	BLEU
1	AALTO	17.37
2	Cmu-cunei	22.38
3	Cu-bojar	19.18
4	Cu-zeman	13.85
5	Google	22.84
6	UEDIN	22.28
7	Comb_eq3	21.89
8	Comb_eq3_sum	22.14
9	Comb_eq4	22.05
10	Comb_eq4_sum	22.04
11	Comb_eq5	22.07

Table 6: Experiment results of linear combination of second set of features

Comb_eq3 or comb_eq4 means that scoring was done using Equation 3 or Equation 4 respectively. Absence of sum means that LM scoring was done using whole sentence as once, while its presence means that log probability of individual trigrams was summed up to get LM score. E.g. Comb_eq4_sum means that final score was calculated using Equation 4 where LM was calculated using sum of log probabilities of trigrams.

The above results are still less than the best performing individual system. But they are much closer to the best system compared to previous experiments. Results showed that having a more lenient feature such as common trigrams rather than NOS of Equation 2 helped in selection of better candidate translations i.e. finding common chunks of words in candidate translations helps.

5.5 Using 3 systems for combination

In order to see the effect of giving fewer choices for candidate selection only 3 best performing systems were used for combination in this series of experiments. It is hoped that lowering the number of choices for candidate selection will help in selection of better hypothesis or candidates and might give some overall improvement.

All previously experimented configurations were used for these experiments with the difference of using lesser number of systems for combination.

Results are presented in the following table.

No.	System	BLEU
1	Cmu-cunei	22.38
2	Google	22.84
3	UEDIN	22.28
4	Comb3_justLM	22.30
5	Comb3_eq2	22.35
6	Comb3_eq3	22.97
7	Comb3_eq3_sum	23.10
8	Comb3_eq4	23.59
9	Comb3_eq4_sum	23.18
10	Comb3_eq5	22.60

Table 7: Combination Using 3 Systems Results

First 3 rows are the BLEU scores of the individual system used for combination. Comb3 means that 3 systems were used for combination. Eq2 and eq4 means that Equation 2 and Equation 4 were used for sentence scoring respectively. Absence or presence of sum serves the same purpose as in previous series of experiments.

Scoring just against LM score and Equation 2 still gave BLEU lower than that of individual systems. Both equations 3 and 4 gave a score higher than that of best performing individual system. BLEU of Equation 5 i.e. common trigrams score without LM score was again lower than best system BLEU.

The above results finally gave some improvement over individual systems showing the potential of sentence level combination techniques. It can be seen from the above results that LM scoring and common trigrams score are useful features for sentence re-ranking. In this series Equation 4 gave better BLEU compared to Equation 5 causing us to conclude that NOW feature is not that useful but experiments of section 5.4 indicated that it cannot always be ignored.

5.6 Comments on Sentence Level Combination

It can be concluded from the experiments that feature selection is the key for sentence level combination techniques though not all configurations were able to beat the best performing system.

Overall increase of 0.75 points in terms of BLEU was achieved in the above experiments. Considering the fact that sentence level combination techniques are quite simple in nature, this improvement is quite encouraging for further investigation of sentence re-ranking techniques for system combination.

One might argue the need for using a higher order language model, which might produce even better results. But this runs counter to the goal of this thesis and also that a higher order LM's increase in complexity precedes its benefits in many cases.

Automatically evaluating a list to get the best sentence goes beyond system combination since it can be used for many other applications. It needs to be studied from many different angles such as evaluating the grammatical and semantic quality of candidate sentences, which on its own is quite a challenging task.

6 Confusion Network Based Experiments

6.1 Introduction

Confusion Network based techniques are the most widely used techniques in system output combination. In general they consists of creating a Confusion Network from the outputs of participating systems and then finding the optimal path in it by scoring against a language model to produce a new translation. Since CN based techniques work on word level they offer a lot of room for improvement in the final translation produced. It is the matter of finding the correct path.

6.2 Confusion Network

A word lattice is defined as a acyclic directed graph with a single starting point and edges labeled with a word or node and it may also has a weight associated with it. A lattice can represent arbitrary mappings between its nodes. Because of this arbitrary mapping a word lattice can represent any finite set of strings.

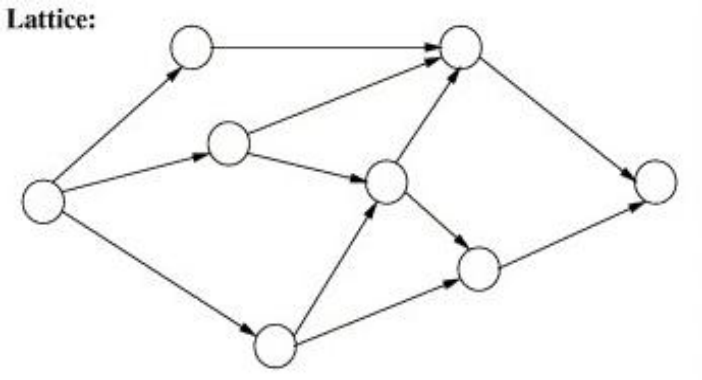


Figure 4 General Diagram of Lattice

Confusion Network can be defined as a compact representation of word lattice. The key difference is that a CN requires that every path must pass through each node, thereby limiting the number of possible paths or in other words limits the number of strings the structure can represent.

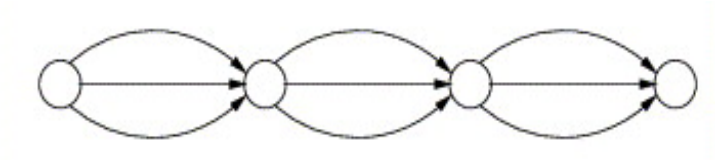


Figure 5 general Digram of Confusion Network

Figure above gives a very general view of a Confusion Network.

A word based Confusion Network can be something like in the figure below.

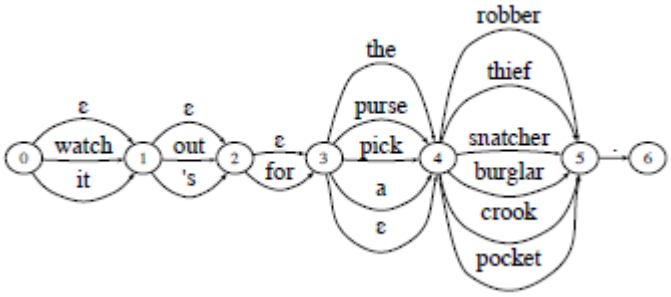


Figure 6 word based Confusion Network

Where each path will denote a word into the final sentence which can be generated by going from node 0 to 6 while choosing the paths with higher probability.

From another point of view a Confusion Network for system output combination can also be viewed as ordered sequence of columns. Each word from each system output corresponds to a particular column. There can also be columns with null entry.

she	bought	the	Jeep	ε
she	buys	the	SUV	ε
she	bought	the	SUV	Jeep

Figure 7 Alternate view of Confusion Network

The job of combination techniques is to find the best path starting from left going to right and producing a new sentence.

6.3 Confusion Network Decoding using Viterbi Decoding

In these experiments CN was created using different schemes but the produced CN was decoded using Viterbi decoding.

6.3.1 Confusion Network Creation using a Backbone

Backbone Selection:

As there can be a big difference between source and target word ordering therefore target translation hypotheses can have varying degree of word ordering. It is important to resort to some particular word ordering in order to get meaningful final translation. It is for this reason there have to be a skeleton or backbone translation to which all other hypotheses will be aligned.

For the purpose of this experiment the best performing system was chosen to be the backbone hypotheses. Although there are many sophisticated method employed by researcher to find the backbone hypotheses (Rosti, et al., 2007), but choosing the best hypothesis as a backbone is also a decent choice. In our case there was no significant difference between the word ordering of different system hypotheses, only difference was that better performing systems were generating longer sentences therefore it was a decent choice to simply select the best performing system as the backbone system.

Alignments:

METEOR tool is basically a word based aligner, which basically does monotonic alignments between outputs and it penalizes cross alignments. It creates alignments as a series of stages, each stage is controlled by a module, where each module perform alignment based upon different criterion e.g. 'exact' module matches only exact stings, while 'stem' uses stems obtained from stemmers to match the two strings and synonym matches strings using its synonym database.

METEOR aligner works with only two sentences at a time (Banerjee, et al., 2005). It aligns the words of candidate string to the words of the reference string. It does so by using above mentioned stages, while each stage is divided into two phases. For the first phase all possible word mappings which can be thought as lines between the words of reference and candidate string, are collected and in the second phase the largest subset of these mappings is selected to produce an alignment somewhat like one shown in figure below.

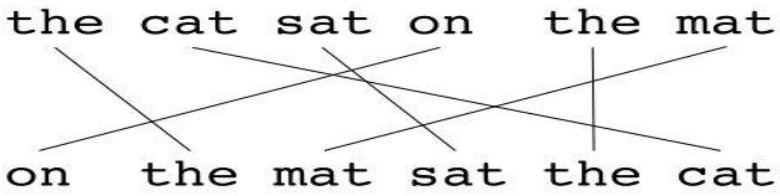


Figure 8 METEOR Alignment Example

Since the experiments in this thesis can use arbitrary number of systems therefore in order to create the Confusion Network, after the selection of backbone each non backbone system output was aligned with the output of backbone system intern and based on those alignments the Confusion Network was created.

CN Decoding

In order to decode the CN for getting the best path i.e. the new generated output. SRI toolkit's lattice-tool was used. Lattice-tool requires that the lattice or in our case the Confusion Network has to be provided in DARPA format or HTK standard lattice format (Young, et al., 2006). Therefore the CN was converted into that format using Hparse tool of HTK toolkit.

Lattice-tool finds the best path in a CN based on probabilities scored against an n-gram language model. Since our previous experiments showed that the LM made from UN and Europarl corpus gave best results therefore for CN based experiments, only LM made from UNEU copra was used.

Results

No.	System	BLEU
1	AALTO	17.37
2	Cmu-cunei	22.38
3	Cu-bojar	19.18
4	Cu-zeman	13.85
5	Google	22.84
6	UEDIN	22.28
7	cn_all6	22.09
8	cn_4	22.20
9	cn_3	22.20
10	cn_6_withempthy	14.15
11	cn_4_withempty	22.20

Table 8 CN experiments with skeleton results

First 6 rows are individual participating system scores presented here for easy comparison. Cn_all6 is the output from the Confusion Network created from all 6 participating systems with output from Google system being the skeleton. Cn_4 is the output from the Confusion Network created from 4 systems with again Google being the skeleton. And cn_3 with being CN created from 3 systems again with the same skeleton. Cn_6_withempty means that all 6 systems were used only difference is that when the skeleton sentence word has no alignment with other system outputs then an empty symbol was inserted in that column, previously nothing was there in that column leaving the CN decoder to choose the skeleton word. But here the decoder can skip that word choosing the empty symbol, which can be removed later before evaluating the output.

None of the experiments were able to beat the best performing system but they all are quite close and reducing the number of systems caused an improvement in the score i.e. 22.09 of 6 systems compared to 22.20 of 3 systems. Similar situation was observed with the introduction of empty columns in the CN. The CN decoder choose the shortest path when given the choice to choose or skip the word in final output, though the score fell a lot when 6 system were used i.e. Cn_6_withempty, but when the number of participating systems was reduced the BLEU score went up again as the number of choices for the CN decoder decreased.

Upon the investigation of the produced output it was found out that the sentences look lot similar to the output of the skeleton systems output. One possible reason for this might be the fact that meteor tries to align the hypotheses as much closely to the reference in our case the skeleton as possible. Most of the words aligned were same in all hypotheses, in other words the skeleton get the lions shares when it comes to decoding the Confusion Network.

6.3.2 CN building without Alignments

Confusion Network Creation

For this series of experiments no alignments were done with any tool to create the Confusion Network. Instead of selecting a skeleton sentence, and then aligning the other systems output to it, to create the CN. The Confusion Network was created simply by putting the individual system outputs parallel to each other i.e. the firs column of the Confusion Network would be the first word of every participating system output. The motivation behind this experiment was to investigate the importance of alignment and importance of reducing the number of options in Confusion Networks, since aligning the outputs to each other reduces the number of paths that can be chosen during decoding.

Decoding

Decoding was done in the same manner as in the previous series, using the lattice-tool of SRILM toolkit and against the trigram LM created from UN and europarl corpus. The results are provided in the table below.

Results

No.	System	BLEU
1	AALTO	17.37
2	Cmu-cunei	22.38
3	Cu-bojar	19.18
4	Cu-zeman	13.85
5	Google	22.84
6	UEDIN	22.28
7	cn_6_noalig	08.99
8	cn_4_noalig	20.82
9	cn_3_noalig	20.90

Table 9 CN without alignment results

First 6 rows are again individual system scores. Cn_6_noalig means all 6 systems were used to create the Confusion Network without using any technique of alignment. And cn_4_noalig and cn_3_noalig means that 4 and 3 systems were used for CN creation respectively.

The results are lower than that of the previous series showing the importance of aligning the outputs. The score drops drastically when 6 systems were used in CN creation without any alignment, but when the number of systems were decreased the BLEU score went up but was still lower than the scores of previous series.

6.3.3 Confusion Network Creation without Skeleton Selection

Confusion Network Creation and Decoding

As it is clear from the previous series that not aligning the system outputs lower the score. Therefore in this series of experiments alignments was done without choosing a skeleton. It was done pretty much the same way as in the experiments with skeleton output i.e. using METEOR aligner.

The difference is that instead of aligning all other outputs to a single system's output different combination were tried. I.e. some systems were aligned to some others systems and then the CN was created by joining those learnt

alignments e.g. system 1 aligned to System 2 and System 3 aligned to System 4 etc. Two ways alignment was also tried e.g. system4 output aligned to system5 output and then system5 output aligned to system4 output after that creating the CN from these learnt alignments.

After the creation of CN, decoding was done in the same manner as in previous experiments.

Results

No.	System	BLEU
1	AALTO	17.37
2	Cmu-cunei	22.38
3	Cu-bojar	19.18
4	Cu-zeman	13.85
5	Google	22.84
6	UEDIN	22.28
7	Cn_noskel_12_34_56	12.26
8	Cn_noskel_26_62_25_52	12.55
9	Cn_noskel_56_25	20.67
10	Cn_noskel_25_52	21.40
11	Cn_noskel_56_65	19.73

Table 10: CN without skeleton results

Rows 1 to 6 represent individual systems. Cn_noskel_12_34_56 means that CN was created by aligning system2 with system1 and system4 with system3 and so on. Row8 and last two rows represent examples of two way alignments e.g. Cn_noskel_56_65 means that CN was created by aligning first system6 by system 5 and then system5 by system6.

All above experiments resulted not only in a BLEU score lower than the best individual system but they are also lower than the results of experiments with skeleton based experiments. It can be concluded that it is better to do alignment against some skeletons.

6.3.4 Upper-bound Experiment

Description

Upper-bound experiments are usually performed to see the maximal results that can be achieved from the combination technique. They show how the decoding procedure limits the output quality here.

Upper-bound experiments for this series involve using the best possible

translation as skeleton in our case the reference and aligning all systems outputs to it and to see how much of the translation is captured by the CN decoder using Confusion Network decoding.

Confusion Network Creation and Decoding

The reference translation is chosen as the skeleton and then all individual system outputs are aligned to it using the METEOR aligner, same way as it was done in the previous series. The CN decoding is also done in the same manner as in the previous series of experiments.

Results

No.	System	BLEU
1	Ul_6	89.89
2	Ul_4	93.12

Table 11 CN Upper-bound Experiments

Ul_6 shows the experiment when all 6 individual systems were aligned to the reference translation with reference being the 7th and backbone system, while Ul_4 shows the experiment result of aligning 4 systems to the reference translation.

The above results shows that the results are very close to the reference translation, showing that impact of better skeleton selection on the performance.

6.4 Confusion Network Decoding using Majority is Authority Decoding

In this experiments series instead of Viterbi decoding using lattice-tool, an alternate method for decoding was implemented and used for decoding the CNs.

6.4.1 Confusion Network Creation

As these experiments only differ in the way the CN is decoded therefore in these experiments all techniques of section 6.3 were used. I.e. CNs were decoded after creating them using all settings of section 6.3 namely skeleton base, non skeleton base and no alignment based CN building. These CNs were then decoded using technique described in next section.

6.4.2 Majority is Authority Decoding

The experiments in Chapter 5 show that common trigrams in candidate translation proved to be a useful feature i.e. commonality in different

candidate translations can be useful for translation improvement. Therefore "Majority is Authority" decoding technique is based on the same idea i.e. to give common elements of candidate translations some importance during combination.

The CN is decoded from left to right as follows. If more than 60% of cells in the column have the same word then that word is selected for the final output. If not then all words of that column are scored against trigram language model using previous two selected words and trigram having the highest probability is selected for the final output.

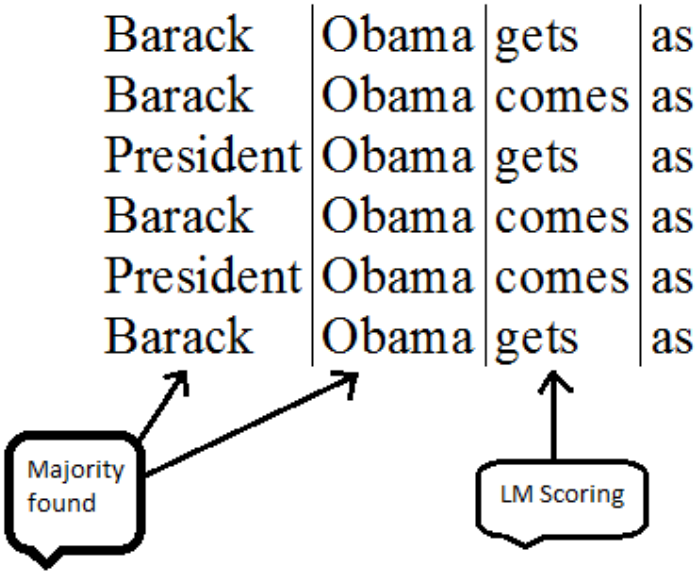


Figure 9: Majority is Authority decoding example

Above diagram shows the decoding process. In the first column 4 out of 6 cells have Barack so it is selected. From the second column Obama is selected. As no word is in majority in the third column therefore both "Barack Obama gets" and "Barack Obama comes" will be scored against trigram LM and the trigram scoring higher will be selected.

For this series of experiments all CNs were decoded using this decoding scheme and final output was produced.

6.4.3 Results

Results of Decoding of CNs by "Majority is Authority" decoding are presented in table below.

No.	System	BLEU
1	AALTO	17.37
2	Cmu-cunei	22.38
3	Cu-bojar	19.18
4	Cu-zeman	13.85
5	Google	22.84
6	UEDIN	22.28
7	Cn_ske_mia_2	22.77
8	Cn_ske_mia_3	22.85
9	Cn_ske_mia_6	22.56
10	Cn_mia_noalig_2	17.33
11	Cn_mia_noalig_3	16.10
12	Cn_mia_noalig_6	10.62
13	Cn_noske_mia_12_34_56	10.57
14	Cn_noske_mia_23_56	15.06
15	Cn_noske_mia_25_56	16.11

Table 12: Majority is Authority Decoding Experiments Results

Rows 1 to 6 are BLEU scores of individual systems. Rows 7 to 9 are skeleton base CNs decoded using MIA decoding. Cn_ske_mia_3 means 3 systems were used to build the CN with one system being the skeleton and it was decoded using MIA. Rows 10 to 12 are CNs created without alignment and Cn_mia_nogalig_6 means that CN was created using 6 systems without aligning them and it was decoded using MIA. Last 3 rows are experiments with CNs made by not selecting a skeleton. Cn_noske_mia_25_56 means that system5 was aligned to system2 and system6 was aligned to system 5 aligned to system6. Long story short CNs for the experiments of rows 13 to 15, were made in a manner similar to section 6.3.3.

Both non skeletons based and no alignment based experiments in addition of scoring less than the best individual system also scored lower than their Viterbi counterparts of section 6.3.2 and 6.3.3. Upon investigation of the produced set it revealed that there was lots of repetition of words in the produced output such as “cooperation among cooperation nation” etc. which were possibly due to the lack or absence of alignment in the CN alternative words. Though a post processing measure of eliminating duplicates such as “Obama Obama says” or “President Barack Obama Barack Obama”was

implemented but some repetitions like the one shown above were not detected leading to a lower BLEU score.

On the other hand skeleton based experiments performed better than all other CN experiments. Not only are they very close to the best individual system but one experiment resulted in BLEU score better than the best individual system i.e. 22.85 of cn_ske_mia_3 compared to 22.84 BLEU of Google giving the only improved translation result in the chapter.

6.5 Comments on Confusion Network Based Experiments

Experiments showed that since Confusion Network based combination works on word level. Therefore they have a lot of margin of creating a different output than just selecting one among the candidates. That output can also be of lower quality i.e. the experiment with 8.99 BLEU score or they can be of good quality.

Experiments also demonstrate the importance of skeleton selection when building CN. The upper-bound experiments showed that choosing the right skeleton really improves the quality since in word to word alignment the skeleton words are more likely to be in the final output. Also the only improved translation result was also with skeleton based CN. Therefore having the right skeleton to do the alignment is the key to get improvement.

As Confusion Network techniques are mostly black box in nature, as it has been shown in the experiments and the literature review. They have a practical applicability in a lot of scenarios since they can be used to combine arbitrary number of systems and all they require is the output of participating systems.

7 Conclusion and Comments

Machine translation systems output combination offers a lot of potential in terms of translation quality improvement. But using very sophisticated and computationally demanding techniques for system combination limits the usability of the approach. Glass box techniques mentioned in the literature review might be valuable from a research point of view but they are not only limiting in the applicability of such approaches but also poses the question whether to perform system combination altogether, since these techniques mean redoing the translation process avoiding weaknesses of one system and utilizing the strength of the other. So it is practical to think that, instead of doing this as an external process it might be better to redesign an individual system.

Getting the outputs from all participating systems including the information regarding their translation process and then using this information to practically making an additional system to redo the translation, leads one with n+1 translation system where n is the number of individual system. This setup is very computationally demanding and is mostly not possible outside a research environment. The most likely usability scenario for using combination techniques is combining output of many available translation systems most of them commercially of the shelf in nature therefore glass box approaches are pretty much out of question for practical usability.

Even with black box techniques, they have to be efficient in terms of resources required. Since in a practical scenario one wants to get the final translation quickly as possible and combining the output is an additional step over getting the output from many different systems therefore it has to be very quick to be practical.

The thesis showed that simple techniques such as sentence re-ranking can be used to improve the translation quality of individual systems. The importance of feature selection was demonstrated and some features for sentence re-ranking were suggested.

The experiments also suggested that general purpose tools for Confusion Network based experiments were not sufficient for improvement advocating

the need for customized path selection techniques more suitable for output combination. The only improved result for CN based experiments came from customized decoding rather than using the general purpose tool.

Efficient techniques for system combination can be valuable tool for achieving improvement in translation quality.

Index

P

R

S

T

U

V

Bibliography

Banerjee Satanjeev and Lavie Alon METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments [Conference] // ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. - Ann Arbor : [s.n.], 2005.

Bangalore Srinivas, Bordel German and Riccardi Giuseppe Computing consensus translation from multiple machine translation systems. [Conference] // IEEE Automatic Speech Recognition and Understanding Workshop. - 2001.

Bojar Ondřej and Žabokrtský Zdeněk CzEng 0.9 Large Parallel Treebank with Rich Annotation [Journal]. - Prague : The Prague Bulletin of Mathematical Linguistics, 2009.

Burch chris, koehn Philipp and Monz Christof Findings of the 2010 JointWorkshop on Statistical Machine Translation and Metrics for Machine Translation [Conference] // Joint 5th Workshop on Statistical Machine Translation and MetricsMATR. - Uppsala : [s.n.], 2010.

Dyer Christopher, Muresan Smaranda and Resnik Philip GeneralizingWord Lattice Translation [Conference] // ACL-08: HLT. - Columbus : [s.n.], 2008.

Feng Yang [et al.] Lattice-based System Combination for Statistical Machine Translation [Conference] // Conference on Empirical Methods in Natural Language Processing. - Singapore : [s.n.], 2009.

Filippova Katja and Strube Michael Tree Linearization in English: Improving Language Model Based Approaches [Conference] // NAACL HLT. - Boulder : [s.n.], 2009.

He Xiaodong and Toutanova Kristina Joint Optimization for Machine Translation System Combination [Conference] // Conference on Empirical Methods in Natural Language Processing. - Singapore : [s.n.], 2009.

Heafield Kenneth and Lavie Alon Combining Machine Translation Output with Open Source The Carnegie Mellon Multi-Engine Machine Translation Scheme [Journal]. - [s.l.] : The Prague Bulletin of Mathematical Linguistics, 2010. - 93.

Heafield Kenneth, Hanneman Greg and Lavie Alon Machine Translation System Combination with FlexibleWord Ordering [Conference] // 4th EACL Workshop on Statistical Machine Translation. - Athens : [s.n.], 2009.

Huang Fei and Papineni Kishore Hierarchical System Combination for Machine Translation [Conference] // Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.. - Prague : [s.n.], 2007.

Hutchins John The first public demonstration of machine translation:the Georgetown-IBM system, 7th January 1954 [Journal]. - 2005.

Karakos Damianos [et al.] Machine Translation System Combination using ITG-based Alignments [Conference] // ACL-08: HLT. - Columbus : [s.n.], 2008.

Khalilov Maxim and Fonollosa José N-gram-based Statistical Machine Translation versus Syntax Augmented Machine Translation: comparison and system combination [Conference] // 12th Conference of the European Chapter of the ACL. - Athens : [s.n.], 2009.

Koehn Philipp Europarl: A Parallel Corpus for Statistical Machine Translation [Conference] // MT summit. - Citeseer : [s.n.], 2005.

Manning Christopher D and Schutze Hinrich Foundations of Statistical Natural Language Processing [Book]. - [s.l.] : MIT Press, 1999.

Matusov Evgeny, Ueffing Nicola and Ney Hermann Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment [Conference] // Cambridge University Engineering Department. - 2006.

Nirenburg Sergei and Frederlcing Robert Toward Multi-Engine Machine Translation [Conference]. - 1994.

Rosti Antti-Veiko [et al.] Combining Outputs from Multiple Machine Translation Systems [Conference] // NAACL HLT. - Rochester : [s.n.], 2007.

Schroeder Josh, Cohn Trevor and Koehn Philipp Word Lattices for Multi-Source Translation [Conference] // Conference of the European Chapter of the ACL. - Athens : [s.n.], 2009.

Sim K C [et al.] CONSENSUS NETWORK DECODING FOR STATISTICAL MACHINE TRANSLATION SYSTEM COMBINATION [Conference] // IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. - 2007.

Specia Lucia [et al.] Estimating the Sentence-Level Quality of Machine Translation Systems [Conference] // 13th Annual Conference of the EAMT. - Barcelona : [s.n.], 2009.

Thurmair Gregor Comparing Rule-based and Statistical MT Output [Conference] // Workshop on the amazing utility of parallel and comparable corpora. - 2004.

Young Steve, Evermann Gunnar and Gales Mark The HTK Book [Book]. - [s.l.] : Cambridge University Engineering Department, 2006.