Domain Adaptation in Statistical Machine Translation

Dimitrios Mavroeidis

Master of Science

Artificial Intelligence

School of Informatics

University of Edinburgh

2007

Abstract

Human beings are capable of categorizing a document based on its topic. Computers are already able to perform very well on that task. However, when translating from one language to another, the human translator will use this knowledge to adapt the writing style and vocabulary for the translation to sound as natural as possible.

Statistical Machine Translation (SMT) uses Probabilistic Machine Learning methods to perform translations. However, such systems do not perform well in domains different from the ones used to train them. How can the ability to recognize the topic of a document be captured by an SMT system to perform better?

Methodologies for adapting a Statistical Machine Translation System to a specific domain are explored. Two methods are examined. The one mixes translation and language models, weighting them appropriately to improve translation quality. The other uses unsupervised methods to cluster a corpus into sub-corpora, train them individually and decode on a specific trained cluster according to the genre or "domain" of the new sentence to be translated.

Experimentation showed improvement in translation quality using both methods. Training on a small domain-specific corpus and a large general one, can improve the performance on translating documents in the small corpus' domain.

Acknowledgements

My heartfelt thanks are to Mr. Josh Shroeder for his eagerness to help throughout the dissertation.

My appreciation to the Arnaoutis Foundation for their financial support.

Lastly, I would like to thank my supervisor Dr. Philipp Koehn for making the dream of Machine Translation seem ever so close.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Dimitrios Mavroeidis)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

When translating from one language to another, a human translator can easily identify the context and domain of the document and use the most appropriate words and style. For instance, when translating Medical documents compared to Law documents, the vocabulary and approach used will be noticeably different.

Word sense disambiguation is a natural outcome of this process. Words like “bat” translate differently based on their meaning (an animal or a sports instrument). Having identified the context or “domain” of the document or sentence in which this word occurred (biology or sports), it is much easier to find the correct translation.

The goal of this dissertation is to enable a Statistical Machine Translation (SMT) System to identify the domain of a document to translate and use the most appropriate vocabulary and style. Until very recently there were no Machine Translation (MT) systems able to perform this kind of identification and use it to improve translation quality.

Two different methods are explored to tackle the problem. The first uses three bilingual corpora (collections of documents) from different sources and of different sizes and domains. We build the models of each corpus separately and try to combine them in different ways to improve translation quality.

The second method uses one bilingual corpus divided into clusters of similar documents. We train on these clusters getting a translation and language model for each one. In the decoding process, a pre-processing step is added in order to identify the domain and use the appropriate models.

A third approach is supposed to combine the two to provide us with even better results. Due to time and infrastructure restrictions, it was not possible to implement.

The structure of this thesis is described in the following lines:

Chapter 2 makes an introduction to Statistical Machine Translation (SMT).

Chapter 3 introduces Domain Adaptation, presents previous work and also describes the data used in this dissertation.

Chapter 4 describes the implementation of the methods and the tools used and code written. Results of experimentation are also presented.

Chapter 5 aggregates results and interpretation into a solid conclusion. Future improvements of the methods used are also discussed here.

Chapter 2

Statistical Machine Translation

2.1 Introduction – Machine Translation

Machine Translation (MT) – also referred to as Automatic Translation – is the mechanization of translation between languages. The primary goal of MT was to create a system capable of taking a natural language text of the source language as input and providing a text in the target language as output. Throughout this document, we will refer to the language whose text is to be translated the source language. The language of the translation will be called the target language.

Researchers have proposed different and often diverse models for MT. A rough categorization of these models is described below:

1. Dictionary-based: straightforward, word-by-word translation, usually without any correlation of meaning between words.
2. Rule-based or Knowledge-based: transforms source text into a language independent representation (interlingua) and then transforms that into the target language. [Lonsdale et al., 1994]
3. Example-based: makes use of bilingual corpora as the knowledge base. [Kay, 1980] [Nagao, 1984]
4. Statistical: uses statistical methods with the help of text corpora combining words or phrases of two languages. [Weaver, 1949] [Brown et al., 1988]

5. Hybrid: a combination of two or more of the above approaches. [Paul et al. 2005]

2.2 Statistical Machine Translation

The statistical approach to Machine Translation was proposed as early as 1949 [Weaver]. The computational complexity of such a task, though, - given the computer systems available at the time - delayed the implementation of this idea by more than 40 years. In 1988 [Brown et al.], in the light of more powerful CPUs and plentiful memory, SMT resurfaced.

The framework, in which SMT functions, consists of at least one bilingual corpus. A model is built from this corpus. This model helps calculate the probability of an element in the source language being equal to another in the target language. This probability can be expressed as follows (using Bayes rule):

$$P(T|S) = \frac{P(T)P(S|T)}{P(S)} \qquad (2.1)$$

The formula above describes the probability that, given a source sentence S in the corpus, a match T from the target language is found.

Nevertheless, we are interested in maximizing the above probability. Thus, formula 2.1 can be rewritten as:

$$\begin{aligned} &argmax_T P(T|S) = \\ &argmax_T \frac{P(T)P(S|T)}{P(S)} = \qquad (2.2) \\ &argmax_T P(T)P(S|T) \end{aligned}$$

$P(T)$ symbolizes the probability provided by the language model. $P(S|T)$ is the probability provided by the translation model. The combination of these two models in that manner is called the **noisy channel model** [Shannon, 1948] [Koehn, 2007].

2.3 IBM Models – Word-based

SMT used to be primarily word-based, meaning that translation is done in a word-for-word manner. We will briefly describe the IBM Models' hierarchy for word-based statistical machine translation. Each consecutive model adds characteristics that enhance complexity, robustness and performance [Brown et al., 1990] [Koehn, 2007]:

Model 1 is the simplest one. It performs pure lexical translation – much like dictionary-based translation. There is no consideration for aligning words between the two corpora. This means that, given the Greek phrase *"ήταν ένα μικρό καράβι"*, the translations "*there was a little boat*" and "*boat was a little there*" are considered to be of the same quality.

Model 2 adds an absolute alignment model. This is represented by a probability distribution $a(i|j, l_T, l_S)$, where i is the position of the target word in the translation and j is the position of the word in the source sentence. l_T and l_S are the length of the target and source sentence respectively.

Model 3 adds a fertility model. Fertility calculates how many words in the target language will be needed to translate one word of the source language (this can be 0, 1, 2 or more words). There is also a case when a word in the target sentence does not have its equivalent in the source sentence. This word is given the NULL tag. Until Model 2, each word in the source language corresponds to another in the target language. There are cases where one word translates into two or more words, or just does not appear in the translation (0 words). For instance, the Greek word “υπάρχει” translates into “there is”.

Model 4 introduces a relative alignment model. A word in the source sentence that is directly linked to a word or words in the target sentence according to the previous models, form a cept. A relative distortion is defined for each output word. Different distributions are defined according to whether the target word is produced by the NULL token, the first word of a cept or subsequent words of a cept.

Model 5 fixes deficiency and some alignment issues. In the previous models, there is no restriction on how many words will be put at the same position in the target sentence. Empty positions in the target sentence are recorded in Model 5, so that each time every new word is put in an empty slot.

Word alignments are built by implementing these models.

2.4 Phrase-based Models

There are cases in translation where one word in the source language is translated in two or more words and vice versa. Word-based models are very weak in dealing with these situations which are not very rare. The concepts behind phrase-based SMT are the same as word-based, but in addition to words, we also consider phrases. A phrase does not need to be of linguistic nature (e.g. noun phrase). In fact, experimentation has shown that linguistically-based phrases degrade the performance of Statistical Translation systems [Yamada and Knight, 2001]. Phrases with more than three words do not have significant impact on performance [Koehn et al., 2003].

The statistical framework is almost identical to the word-based models. A new factor ω is introduced which is used to control the output length. This, along with the language model, improves performance greatly.

$$argmax_T P(T|S) = argmax_T P(T)P(S|T)\omega^{length(T)} \tag{2.3}$$

The decoder is the program that finds the best translation by calculating appropriate probabilities and producing the sentence that yields the greatest one. This problem is NP-complete, because word reordering is allowed [Knight 1999]. The dominant approaches for this problem usually lie in the areas of dynamic programming and greedy algorithms [Knight and Marcu, 2004]. Moses, the SMT system used in the dissertation, uses a beam-search algorithm, similar to the one used widely in speech recognition [Jelinek, 1997].

Word-based models provide a good word alignment between each sentence pair in the parallel corpus. In order to use phrase-based SMT effectively, a good phrase translation phrase table should exist. This can be built by extracting phrase pairs from the word alignments. These phrase pairs need to be consistent with the alignment. What is meant with consistent?

A phrase pair$(\bar{T},\bar{S})$ is said to be consistent with an alignment A if all words $s_1,\dots,s_n$ in $\bar{S}$ that have alignment points in A have these with words $t_1,\dots,t_n$ in $\bar{T}$ and vice versa. The definition is given below:

$$\begin{gathered}\forall t_i \in \bar{T}: (t_i, s_j) \in A \rightarrow s_j \in \bar{S} \\ AND\ \forall s_i \in \bar{S}: (t_i, s_j) \in A \rightarrow t_j \\ AND\ \exists s_i \in \bar{S}, t_i \in \bar{T}: (t_i, s_j) \in A\end{gathered} \tag{2.4}$$

This tells us that unaligned words do not violate consistency. They can appear anywhere in the sentence. Nevertheless, at least one alignment point should exist per phrase pair (last line).

Now we have to extract all phrase pairs from each sentence pair that conform to the definition above. The less word alignments there are, the more phrase pairs can be extracted [Koehn, 2007].

In order to build the phrase translation table we also need the probabilities of each phrase pair. The way the table is built is different from what was discussed in the IBM Models. That is because we do not want to eliminate any small phrase due to scarceness. Every phrase pair with its probability can be useful. The following procedure is followed:

A number of phrase pairs is defined for each sentence pair, as described above. The number of times a phrase pair occurs in all the sentence pairs ($count(\bar{T},\bar{S})$ is then divided with all the occurrences of all phrase pairs $\sum_{\bar{S}_i} count(\bar{T},\bar{S}_i)$. Thus, the translation probability is calculated by:

$$\varphi(\bar{S}|\bar{T}) = \frac{count(\bar{T},\bar{S})}{\sum_{\bar{S}_i} count(\bar{T},\bar{S}_i)} \tag{2.5}$$

Even the simplest form of the phrase-based model outperforms the word-based one. Several improvements of the phrase-based model rise performance to even higher levels. So far, we had three ingredients in the model: the phrase table, the reordering model and the language model. Adding weights to these three components, a log-linear model is introduced. It is not always correct to assume that all three components should count the same in the translation model. So, weights are learned in order to give each the importance needed. Naïve Bayes and maximum entropy are the most common log-linear-based machine learning methods. Generally, turning to a log-linear model can improve translation quality, giving – for instance – greater weight to the language model.

Log-linear models have another advantage which is not immediately apparent. Other model components can be added without difficulty. Some components that have been shown to improve translation quality are briefly described below:

- Lexical weighting. Trying to figure out how reliable a rare phrase pair is. If it is, then a greater weight will be given to it to improve performance.
- Bidirectional training. Training the model in both directions – source to target and target to source. Using appropriate weights, it usually outperforms unidirectional models.

2.5 Moses SMT System

The "Moses" Statistical Machine Translation system was developed at the University of Edinburgh. It is based on the phrase-based translation model [Koehn et al., 2003]. "Moses" is a combination of several natural language and machine learning tools.

Input:

- Bilingual parallel corpus. An aligned bilingual corpus must be fed to Moses.
- Language model data (if different from training data).
- Test data. Some input and respective reference files.
- Tuning data. One input and one reference file to help in tuning.
- Evaluation data. Usually the same as the test data.

Process:

- Giza++

- Translation Model Training
- Language Model Training
- Tuning
- Testing
- Evaluation

By-products:

- Language Model
- Translation Model
- Configuration weights

Output:

- Translations of the test data
- BLUE (or any other metric) scores of the test data

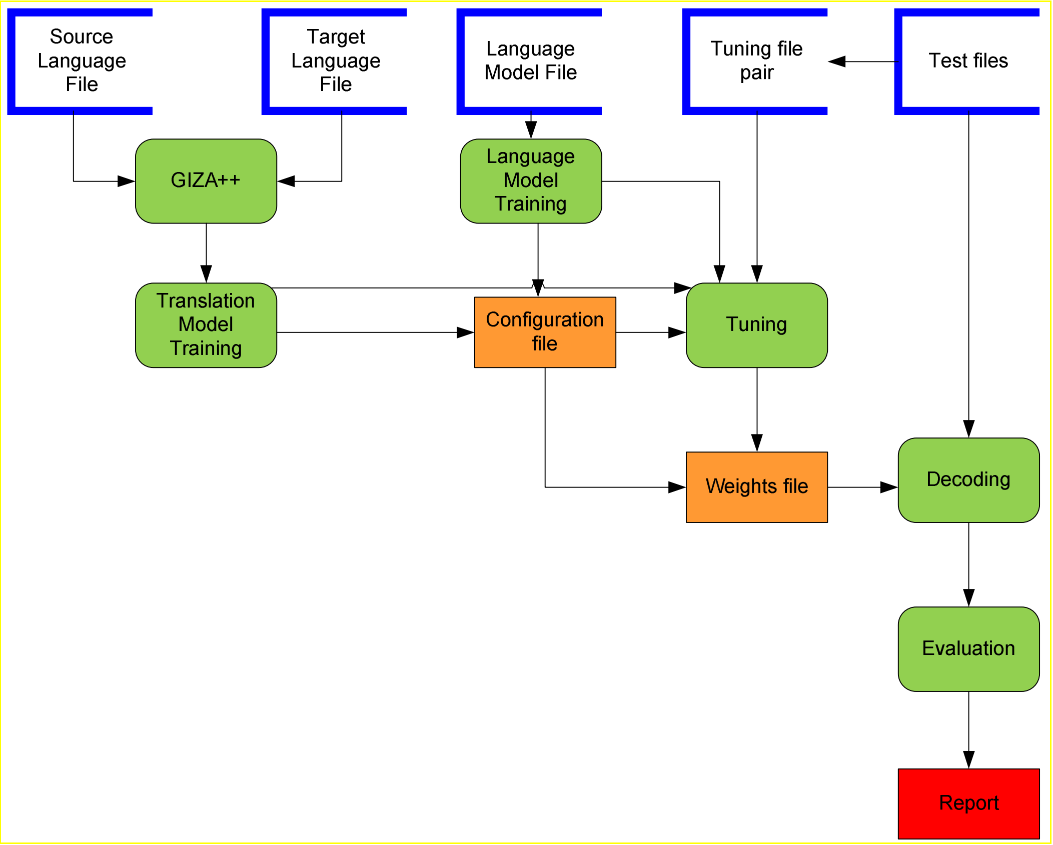


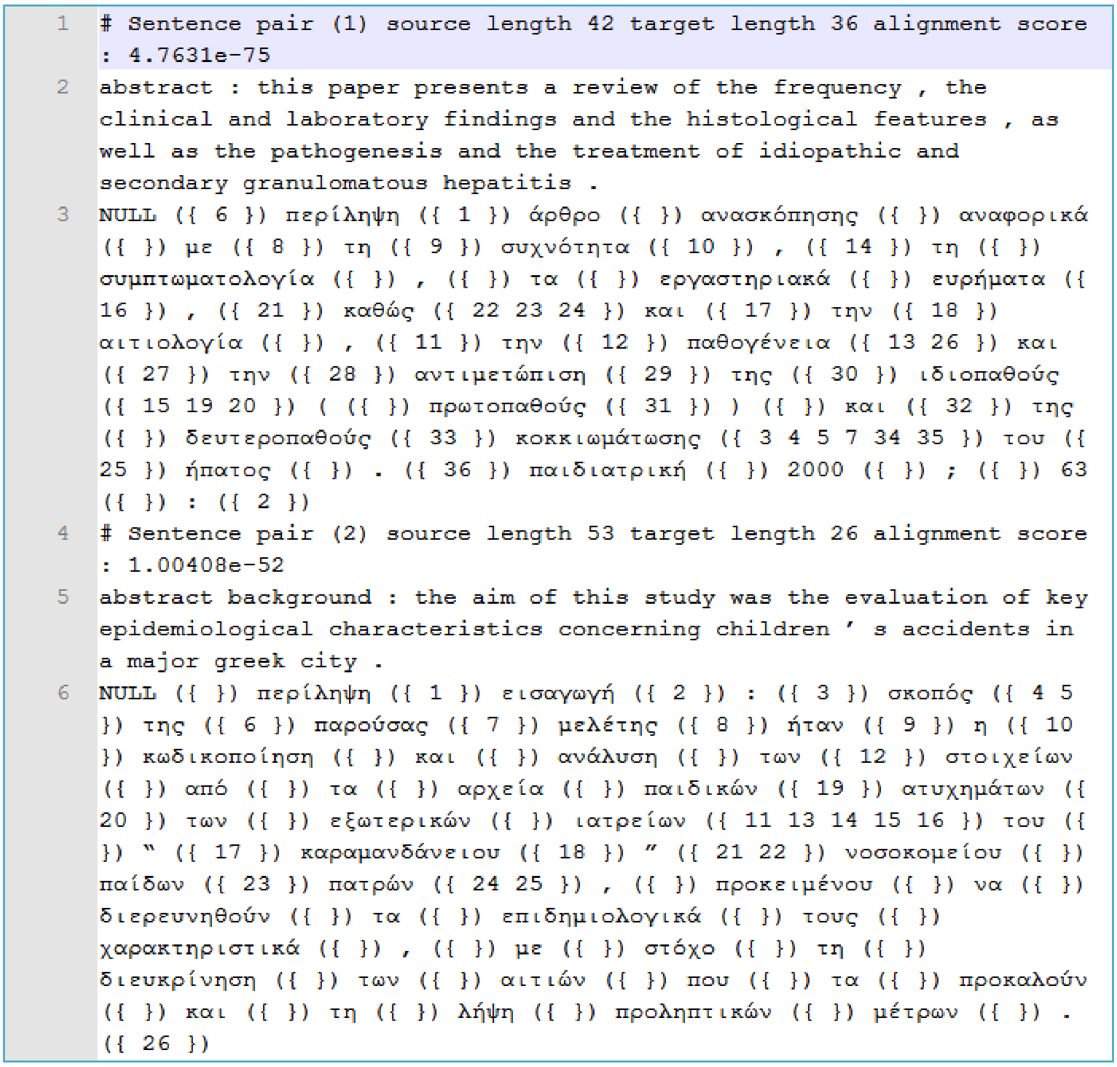
Figure 1: A simplified representation of the "Moses" Statistical Translation System

Bellow, we will describe the most important components of Moses.

2.5.1 Building the word alignment (GIZA++)

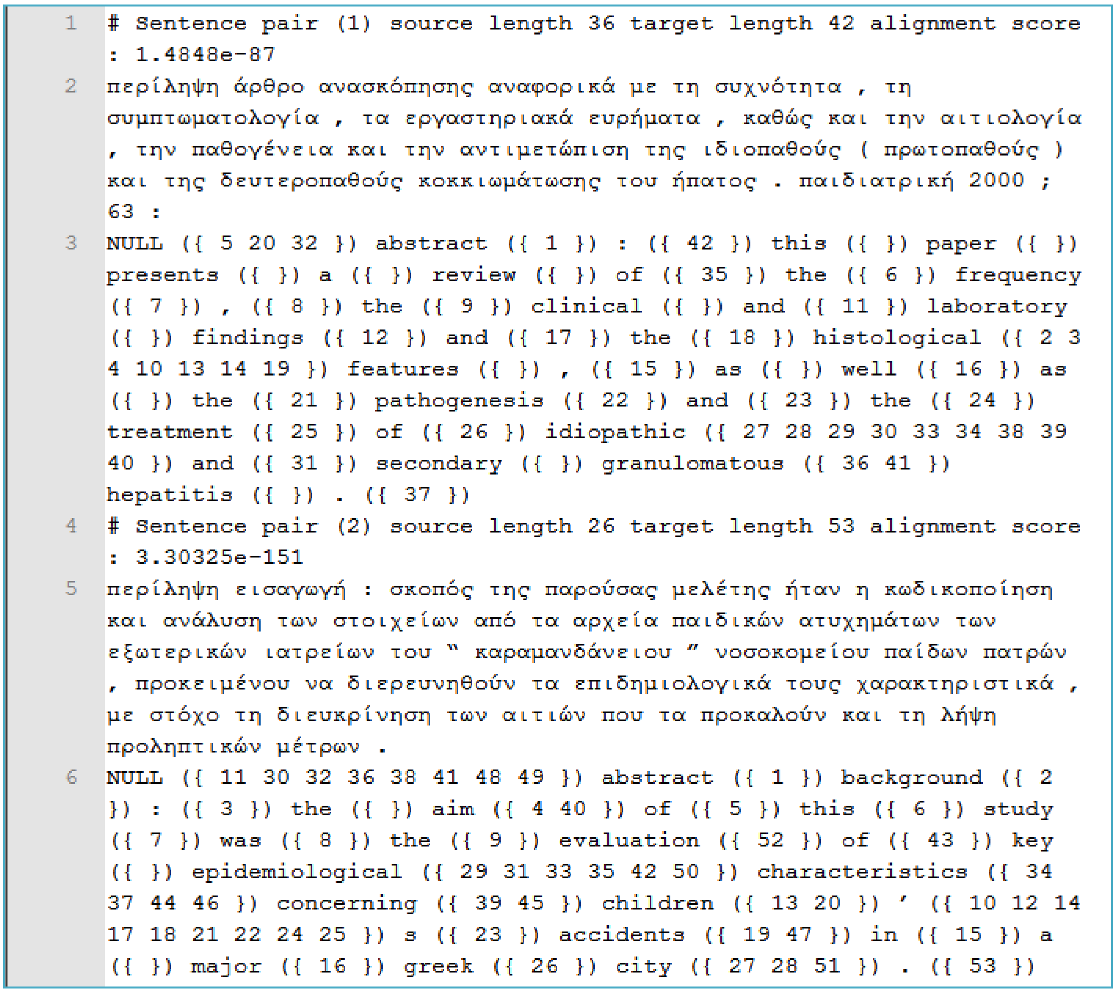
GIZA++ is the tool used to train on the bilingual corpus and build the translation and reordering tables. *GIZA++* is an extension of the program GIZA, developed by the Statistical Machine Translation team during a summer workshop in 1999 at the Center for Language and Speech Processing at Johns-Hopkins University (CLSP/JHU), USA. This extension was designed and implemented by Franz Josef Och [Och and Ney, 2003].

It is used as an initial step to establish word alignments. The word alignments are taken from the intersection of two runs as explained in 2.4. Some additional alignment points are taken from the union of these runs. Only the word alignment part of the IBM Models is required here. The phrase-based model will be built based on the word alignments.



```
# Sentence pair (1) source length 42 target length 36 alignment score
: 4.7631e-75
abstract : this paper presents a review of the frequency , the
clinical and laboratory findings and the histological features , as
well as the pathogenesis and the treatment of idiopathic and
secondary granulomatous hepatitis .
NULL ({ 6 }) περίληψη ({ 1 }) άρθρο ({ }) ανασκόπησης ({ }) αναφορικά
({ }) με ({ 8 }) τη ({ 9 }) συχνότητα ({ 10 }) , ({ 14 }) τη ({ })
συμπτωματολογία ({ }) , ({ }) τα ({ }) εργαστηριακά ({ }) ευρήματα ({
16 }) , ({ 21 }) καθώς ({ 22 23 24 }) και ({ 17 }) την ({ 18 })
αιτιολογία ({ }) , ({ 11 }) την ({ 12 }) παθογένεια ({ 13 26 }) και
({ 27 }) την ({ 28 }) αντιμετώπιση ({ 29 }) της ({ 30 }) ιδιοπαθούς
({ 15 19 20 }) ( ({ }) πρωτοπαθούς ({ 31 }) ) ({ }) και ({ 32 }) της
({ }) δευτεροπαθούς ({ 33 }) κοκκιωμάτωσης ({ 3 4 5 7 34 35 }) του ({
25 }) ήπατος ({ }) . ({ 36 }) παιδιατρική ({ }) 2000 ({ }) ; ({ }) 63
({ }) : ({ 2 })
# Sentence pair (2) source length 53 target length 26 alignment score
: 1.00408e-52
abstract background : the aim of this study was the evaluation of key
epidemiological characteristics concerning children ’ s accidents in
a major greek city .
NULL ({ }) περίληψη ({ 1 }) εισαγωγή ({ 2 }) : ({ 3 }) σκοπός ({ 4 5
}) της ({ 6 }) παρούσας ({ 7 }) μελέτης ({ 8 }) ήταν ({ 9 }) η ({ 10
}) κωδικοποίηση ({ }) και ({ }) ανάλυση ({ }) των ({ 12 }) στοιχείων
({ }) από ({ }) τα ({ }) αρχεία ({ }) παιδικών ({ 19 }) ατυχημάτων ({
20 }) των ({ }) εξωτερικών ({ }) ιατρείων ({ 11 13 14 15 16 }) του ({
}) “ ({ 17 }) καραμανδάνειου ({ 18 }) ” ({ 21 22 }) νοσοκομείου ({ })
παίδων ({ 23 }) πατρών ({ 24 25 }) , ({ }) προκειμένου ({ }) να ({ })
διερευνηθούν ({ }) τα ({ }) επιδημιολογικά ({ }) τους ({ })
χαρακτηριστικά ({ }) , ({ }) με ({ }) στόχο ({ }) τη ({ })
διευκρίνηση ({ }) των ({ }) αιτιών ({ }) που ({ }) τα ({ }) προκαλούν
({ }) και ({ }) τη ({ }) λήψη ({ }) προληπτικών ({ }) μέτρων ({ }) .
({ 26 })
```

Figure 2: The first few lines of a Giza++ word-aligned file from a direct run, trained on an English-to-Greek Medical bilingual corpus



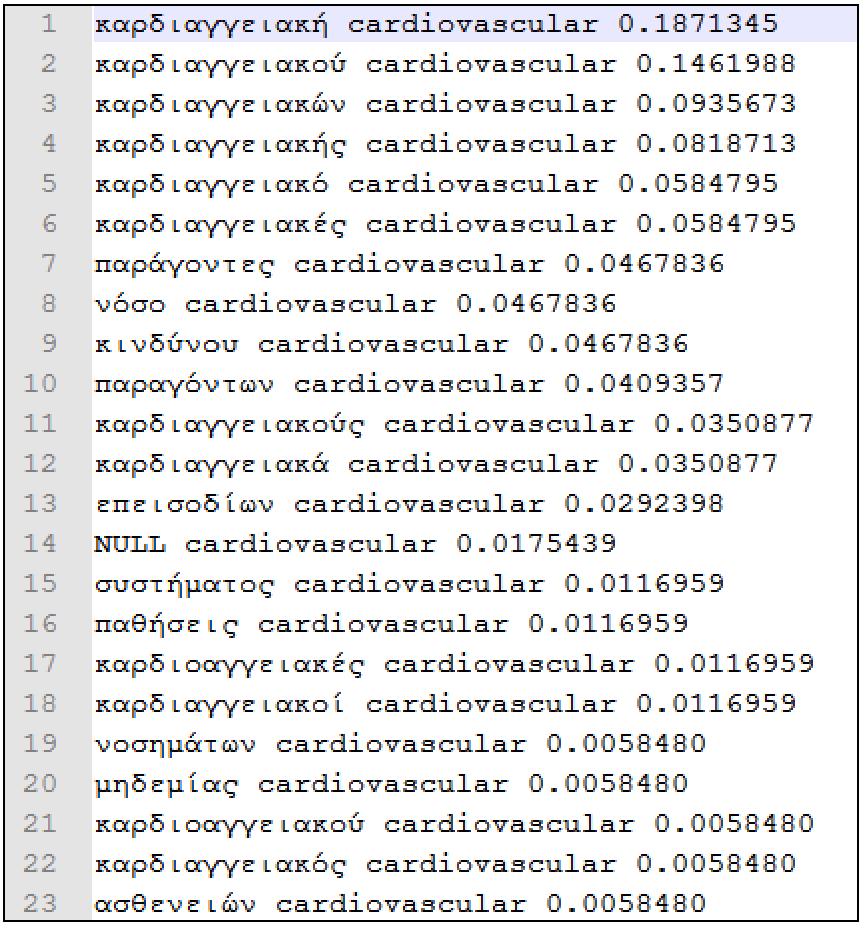
```
# Sentence pair (1) source length 36 target length 42 alignment score : 1.4848e-87
περίληψη άρθρο ανασκόπησης αναφορικά με τη συχνότητα , τη συμπτωματολογία , τα εργαστηριακά ευρήματα , καθώς και την αιτιολογία , την παθογένεια και την αντιμετώπιση της ιδιοπαθούς ( πρωτοπαθούς ) και της δευτεροπαθούς κοκκιωμάτωσης του ήπατος . παιδιατρική 2000 ; 63 :
NULL ({ 5 20 32 }) abstract ({ 1 }) : ({ 42 }) this ({ }) paper ({ }) presents ({ }) a ({ }) review ({ }) of ({ 35 }) the ({ 6 }) frequency ({ 7 }) , ({ 8 }) the ({ 9 }) clinical ({ }) and ({ 11 }) laboratory ({ }) findings ({ 12 }) and ({ 17 }) the ({ 18 }) histological ({ 2 3 4 10 13 14 19 }) features ({ }) , ({ 15 }) as ({ }) well ({ 16 }) as ({ }) the ({ 21 }) pathogenesis ({ 22 }) and ({ 23 }) the ({ 24 }) treatment ({ 25 }) of ({ 26 }) idiopathic ({ 27 28 29 30 33 34 38 39 40 }) and ({ 31 }) secondary ({ }) granulomatous ({ 36 41 }) hepatitis ({ }) . ({ 37 })
# Sentence pair (2) source length 26 target length 53 alignment score : 3.30325e-151
περίληψη εισαγωγή : σκοπός της παρούσας μελέτης ήταν η κωδικοποίηση και ανάλυση των στοιχείων από τα αρχεία παιδικών ατυχημάτων των εξωτερικών ιατρείων του " καραμανδάνειου " νοσοκομείου παίδων πατρών , προκειμένου να διερευνηθούν τα επιδημιολογικά τους χαρακτηριστικά , με στόχο τη διευκρίνηση των αιτιών που τα προκαλούν και τη λήψη προληπτικών μέτρων .
NULL ({ 11 30 32 36 38 41 48 49 }) abstract ({ 1 }) background ({ 2 }) : ({ 3 }) the ({ }) aim ({ 4 40 }) of ({ 5 }) this ({ 6 }) study ({ 7 }) was ({ 8 }) the ({ 9 }) evaluation ({ 52 }) of ({ 43 }) key ({ }) epidemiological ({ 29 31 33 35 42 50 }) characteristics ({ 34 37 44 46 }) concerning ({ 39 45 }) children ({ 13 20 }) ' ({ 10 12 14 17 18 21 22 24 25 }) s ({ 23 }) accidents ({ 19 47 }) in ({ 15 }) a ({ }) major ({ 16 }) greek ({ 26 }) city ({ 27 28 51 }) . ({ 53 })
```

Figure 3: The first few lines of a Giza++ word aligned file from an inverse run, trained on an English-to-Greek Medical bilingual corpus

GIZA++ requires a large amount of memory (typically more than 1 GB). There are options that enable GIZA++ to run training in parts in order to avoid the memory problem.

After GIZA++, a program named “grow-diag-final” is run intersecting the two runs of GIZA++ and adding some extra alignment points (as described in 2.4).

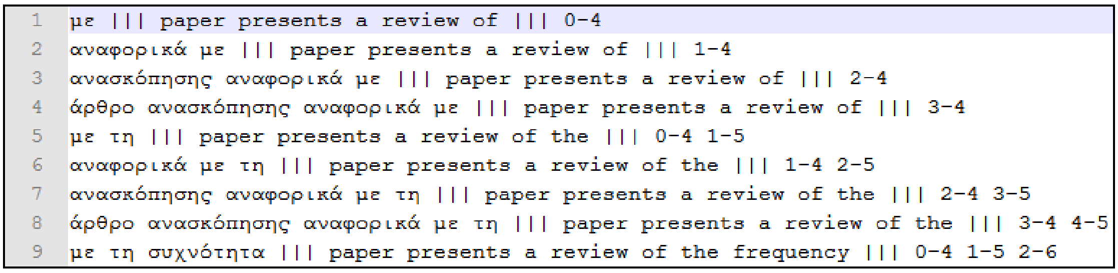
Building the phrase translation table follows. The first step is to calculate the likelihood of word pairs (target-source). This is quite trivial. In Figure 4, we present the maximum likelihood estimation for the word “cardiovascular” from the same corpus.



```
καρδιαγγειακή cardiovascular 0.1871345
καρδιαγγειακού cardiovascular 0.1461988
καρδιαγγειακών cardiovascular 0.0935673
καρδιαγγειακής cardiovascular 0.0818713
καρδιαγγειακό cardiovascular 0.0584795
καρδιαγγειακές cardiovascular 0.0584795
παράγοντες cardiovascular 0.0467836
νόσο cardiovascular 0.0467836
κινδύνου cardiovascular 0.0467836
παραγόντων cardiovascular 0.0409357
καρδιαγγειακούς cardiovascular 0.0350877
καρδιαγγειακά cardiovascular 0.0350877
επεισοδίων cardiovascular 0.0292398
NULL cardiovascular 0.0175439
συστήματος cardiovascular 0.0116959
παθήσεις cardiovascular 0.0116959
καρδιοαγγειακές cardiovascular 0.0116959
καρδιαγγειακοί cardiovascular 0.0116959
νοσημάτων cardiovascular 0.0058480
μηδεμίας cardiovascular 0.0058480
καρδιοαγγειακού cardiovascular 0.0058480
καρδιαγγειακός cardiovascular 0.0058480
ασθενειών cardiovascular 0.0058480
```

Figure 4: Lexical probabilities of all the ways we can get the word "cardiovascular" in English from Greek, trained on the Medical bilingual corpus, ceteris paribus.

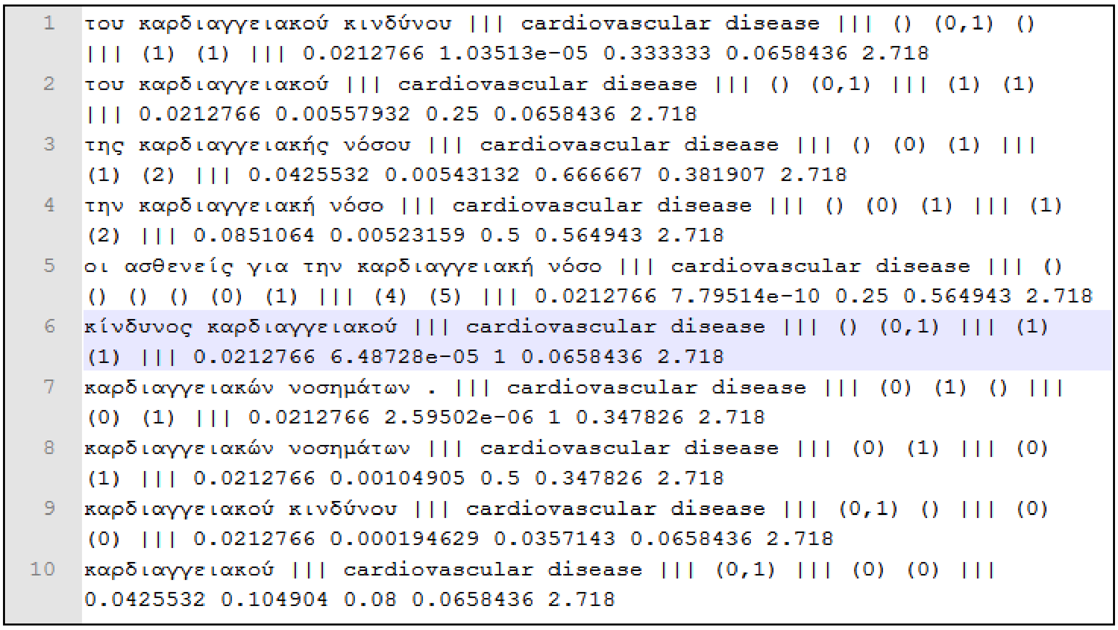
The second step includes phrase extraction. Each source phrase is aligned with its counterpart in the target language. The word alignments are denoted in the form "S-T" where "S" is the position of the word in the source language and "T" the position of the translation in the target language. Figure 5 shows the first few lines of such a file.



```
με ||| paper presents a review of ||| 0-4
αναφορικά με ||| paper presents a review of ||| 1-4
ανασκόπησης αναφορικά με ||| paper presents a review of ||| 2-4
άρθρο ανασκόπησης αναφορικά με ||| paper presents a review of ||| 3-4
με τη ||| paper presents a review of the ||| 0-4 1-5
αναφορικά με τη ||| paper presents a review of the ||| 1-4 2-5
ανασκόπησης αναφορικά με τη ||| paper presents a review of the ||| 2-4 3-5
άρθρο ανασκόπησης αναφορικά με τη ||| paper presents a review of the ||| 3-4 4-5
με τη συχνότητα ||| paper presents a review of the frequency ||| 0-4 1-5 2-6
```

Figure 5: Aligned phrases with the words alignments denoted as a pair of numbers (positions) in the phrase.

Apparently, that is all that is needed to build the phrase translation table. The previous step was necessary in order to avoid loading the whole of the table in memory. Having the aligned-phrases' files and the lexical-probabilities file, the phrase translation table can be built. The phrase translation probability is calculated using formula (2.5). First, the aligned-phrases file is sorted so that all target phrases corresponding to a foreign phrase are the one under the other. In that way, one source phrase can be processed at each one time, all counts for that phrase merged and $\varphi(\bar{S}|\bar{T})$ computed. Estimating the inverse probability $\varphi(\bar{T}|\bar{S})$ is done in the same way, but this time the inverse file is sorted and processed. The numbers seen at the right of the two phrases represent the phrase translation probabilities (both direct and inverse), the lexical weights (direct and inverse) and a phrase penalty. Figure 6 shows part of a phrase table.



```
του καρδιαγγειακού κινδύνου ||| cardiovascular disease ||| () (0,1) ()
||| (1) (1) ||| 0.0212766 1.03513e-05 0.333333 0.0658436 2.718
του καρδιαγγειακού ||| cardiovascular disease ||| () (0,1) ||| (1) (1)
||| 0.0212766 0.00557932 0.25 0.0658436 2.718
της καρδιαγγειακής νόσου ||| cardiovascular disease ||| () (0) (1) |||
(1) (2) ||| 0.0425532 0.00543132 0.666667 0.381907 2.718
την καρδιαγγειακή νόσο ||| cardiovascular disease ||| () (0) (1) ||| (1)
(2) ||| 0.0851064 0.00523159 0.5 0.564943 2.718
οι ασθενείς για την καρδιαγγειακή νόσο ||| cardiovascular disease ||| ()
() () () (0) (1) ||| (4) (5) ||| 0.0212766 7.79514e-10 0.25 0.564943 2.718
κίνδυνος καρδιαγγειακού ||| cardiovascular disease ||| () (0,1) ||| (1)
(1) ||| 0.0212766 6.48728e-05 1 0.0658436 2.718
καρδιαγγειακών νοσημάτων . ||| cardiovascular disease ||| (0) (1) () |||
(0) (1) ||| 0.0212766 2.59502e-06 1 0.347826 2.718
καρδιαγγειακών νοσημάτων ||| cardiovascular disease ||| (0) (1) ||| (0)
(1) ||| 0.0212766 0.00104905 0.5 0.347826 2.718
καρδιαγγειακού κινδύνου ||| cardiovascular disease ||| (0,1) () ||| (0)
(0) ||| 0.0212766 0.000194629 0.0357143 0.0658436 2.718
καρδιαγγειακού ||| cardiovascular disease ||| (0,1) ||| (0) (0) |||
0.0425532 0.104904 0.08 0.0658436 2.718
```

Figure 6: Part of the Greek-to-English phrase table for the Medical bilingual corpus (phrases that translate to "cardiovascular disease").

The final step of training is producing a configuration file that contains all the paths with the trained files and some parameter settings for the decoder.

2.5.2 Language Model Training

As it is apparent from the definition of a translation model (formulas 2.2-2.3), a language model is needed to improve translation quality. A language model contains a list of n-grams that are given a probability according to the frequency they appear in the corpus it is trained on. LMs are often used in applications like speech recognition, part-of-speech tagging, spell-checking, etc. It has also been proven very useful to check if a produced translation sounds natural in the target language.

Translation models are currently insufficient for capturing the style of language. Naturalness in the language produced by any translation system is one of the top priorities. When adding a language model in the process, it intervenes during decoding in order to reinforce more naturally sounding word sequences. Thus, when the system has to choose between two translated sentences that have almost identical probability score according to the phrase translation table, it will choose the one that sounds more natural in the target language (according to the LM).

Language models need not train on the same text as the bilingual corpus. Any text of good target language quality can be added to the model, as only monolingual corpora are needed. The use of language models in the decoding process have a positive effect on the quality of produced sentences, especially in language naturalness. The higher the order of the language model (i.e. using 5-grams instead of 3-grams) improves the quality further. Nevertheless, LM order must not surpass a certain limit, as it makes tuning and decoding rather time-consuming and memory inefficient. In general, the order of the language model has a positive impact on translation quality.

Both SRILM [Stolcke, 2002] and IRSTLM [Federico and Cettolo, 2007] can be used with MOSES.

2.5.3 Tuning

The weights described in the construction of the phrase translation table above are usually not optimal. The process of tuning optimises these weights based on a pair of tuning files. The program that tunes the model is called MERT. This pair is usually taken from the test data. The first step in tuning is to run the decoder on the source tuning file and score it with a metric (BLUE or other). Then, MERT changes the weights according to a maximization algorithm and runs the decoder again. If the quality of translation is improved, then the weights are kept unless a subsequent run improves them better. Finally, the weights that give the best score are stored and used in testing and evaluation.

2.5.4 Decoding

The decoder uses a heuristic algorithm for finding the best translation. It is called beam-search and is an improvement of the best-first search algorithm. It unfolds only the *m* first nodes at each depth. *m* is a constant number. The larger *m* is, the more beam-search resembles best-first search.

The decoder consults the phrase tables created in training and translates each input sentence from the test files. How the phrase table is created, was discussed earlier. The best translation is chosen by a log-linear model that maximizes the following:

$$score(T|S) = exp \sum_{i} \lambda_i h_i(T,S) \qquad (2.6)$$

The weights are created in the manner we described previously in this chapter. The components used are the direct and inverse translation probabilities, the direct and inverse word probabilities, language model, lexicalized reordering model, and phrase and word counts.

2.5.5 Evaluation

For many years, it has been believed that evaluating translations produced by a MT system or even from a human translator is a matter of subjective judgment and should always be done by humans. In 2001, it was shown that this assumption is not very accurate [Papineni et al. 2001]. An algorithm was developed to determine the quality of a translation by counting the number of n-grams that co-occur in the translation and a set of reference translations suggesting the BLEU evaluation system. The BLEU system is now the dominant evaluator for MT systems at NIST and many other organizations. It is also the one that is used more often in Moses than any other metric. Translations in this thesis are evaluated by BLEU.

Chapter 3

Domain Adaptation in SMT

3.1. Introduction

Domain adaptation is defined as the ability of an SMT system to translate successfully any input sentence or document regardless of its genre or "domain". This means that either the input document or sentence is a medical or law one, the translation system should produce an impeccable translation.

Until recently, SMT systems only considered a very limited local context (a few surrounding words) to make disambiguation decisions and did not consider more general properties such as context.

3.2. Previous Work

This year's task for the Workshop on SMT (WSMT 2007) was to use a large out-of-domain training dataset and a much smaller (forty times smaller) in-domain dataset to maximize translation quality in that domain. Two papers from the SMTW 2007 were immediately distinguished due to the similarity with the ideas followed in this thesis.

The first is titled "Experiments in Domain Adaptation for Statistical Machine Translation" [Koehn and Schroeder, 2007]. The methodology followed includes merging different domains' translation and language models in various ways to improve translation quality. The "Moses" SMT system is used.

The other methodology tried was inspired by “Bilingual Cluster Based Models for Statistical Machine Translation” [H. Yamamoto and E. Sumita, 2007]. They divide a corpus into groups of similar documents using an unsupervised method (clustering). They train each cluster individually and during decoding they choose the translation and language model to use by categorizing the new sentence in one of the clusters created before training. Clustering of the data is conducted on a sentence-by-sentence basis.

Another approach worth mentioning is titled “Domain Adaptation in Statistical Machine Translation with Mixture Modelling [J. Civera and A. Juan, 2007]. A mixture extension of Hidden Markov Model (HMM) alignment model and the derivation of Viterbi alignments is used.

This dissertation tries to confirm both of the approaches mentioned above. It also combines them to achieve even better results. Three datasets (bilingual corpora) are used. The first one is Europarl which has been extensively tested on “Moses”. The other two are the JRC-Acquis and a hand-made Medical corpus taken from several “pdf” and “doc” files. The last two have not been tried before on “Moses”. For all three corpora, we use the Greek-to-English language pair.

3.3. Data

As mentioned earlier, large bilingual corpora need to be available in order to train an SMT system. Those corpora need to explicitly associate a word/phrase/sentence/paragraph of the source language to its translation in the target language. This is the concept of alignment. Correct alignment is a crucial factor to the performance of any SMT system.

“Moses” requires that the bilingual corpus consists of two files, one for the source and one for the target language. These are simple text files that contain no other information than simple text (no XML, SGML etc.). Each line in the source language file must correspond to its translation in the target language file. Table 1 shows an example of such an alignment.

Line Number	English Text File	Greek Text File
1	COMMISSION REGULATION (EEC) No 3812/85	ΚΑΝΟΝΙΣΜΟΣ (ΕΟΚ) αριθ. 3812/85 ΤΗΣ ΕΠΙΤΡΟΠΗΣ
2	of 20 December 1985	της 20ής Δεκεμβρίου 1985
3	adjusting certain Regulations on milk and milk products by reason of the accession of Spain	για την αναπροσαρμογή ορισμένων κανονισμών στον τομέα του γάλακτος και των γαλακτοκομικών προϊόντων, λόγω της προσχώρησης της Ισπανίας
4	THE COMMISSION OF THE EUROPEAN COMMUNITIES,	Η ΕΠΙΤΡΟΠΗ ΤΩΝ ΕΥΡΩΠΑΪΚΩΝ ΚΟΙΝΟΤΗΤΩΝ,
5	Having regard to the Treaty establishing the European Economic Community,	Έχοντας υπόψη: τη συνθήκη για την ίδρυση της Ευρωπαϊκής Οικονομικής Κοινότητας,
6	Having regard to the Act of Accession of Spain and Portugal (1), and in particular Article 396 thereof,	την πράξη προσχώρησης της Ισπανίας και της Πορτογαλίας (1), και ιδίως το άρθρο 396, Εκτιμώντας:

Table 1: Part of an aligned corpus (taken from the JRC-Acquis parallel Greek-English

3.3.1 Preparing the Data

For the purpose of this project, three corpora were used:

1. Europarl – extracted from the proceedings of the European Parliament[1]. It includes versions in 11 European languages: Romanic

[1] http://www3.europarl.eu.int/

(French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish.

2. JRC-Acquis – the total body of European Union (EU) law applicable in the the EU Member States. Version 2.2 is used as version 3 only became available in mid-July 2007. The difference between the two is that the latter also contains the legal documents of 2005 and 2006.
3. Medical – A corpus of medical abstracts from the Greek Journal of Paediatrics.

For each of the three corpora, the Greek- English pair was used. As previously described, we need two files for each corpus. In the following paragraphs, the process of pre-processing the data for each corpus will be described:

3.3.2 Europarl Corpus

Data from Europarl were ready to process as they are used extensively by the Edinburgh Machine Translation group to conduct their experiments. The latest version of Europarl (v.3) was compiled using the UTF-8 character encoding. The ISO-8859-XX prototype uses 8 bits per character. This enables the representation of only 256 characters per character set, which means that only 2 languages at most can be represented. Usually, the first 128 ASCII codes represent the Latin alphabet, while the rest represent the characters of another language. UTF-8 uses 16 bits per character, which can represent 65536 characters. Consequently, this encoding can include all known alphabets – even the very rich ones like the Chinese. This was done in order to have a uniform representation of text and not having to change character encoding each time a different language was used for training, adding to the flexibility of Europarl. The downside is that files are twice as big and so are memory demands while processing them. Both the JRC-Acquis and the

Medical corpora were compiled in order to conform to UTF-8 for comparison and combination reasons.

	Characters	Words	Paragraphs
English	89,140,279	15,031,867	536,318
Greek	99,053,585	14,954,294	

Table 2: The Greek-English Europarl corpus

3.3.3 JRC-Acquis Corpus

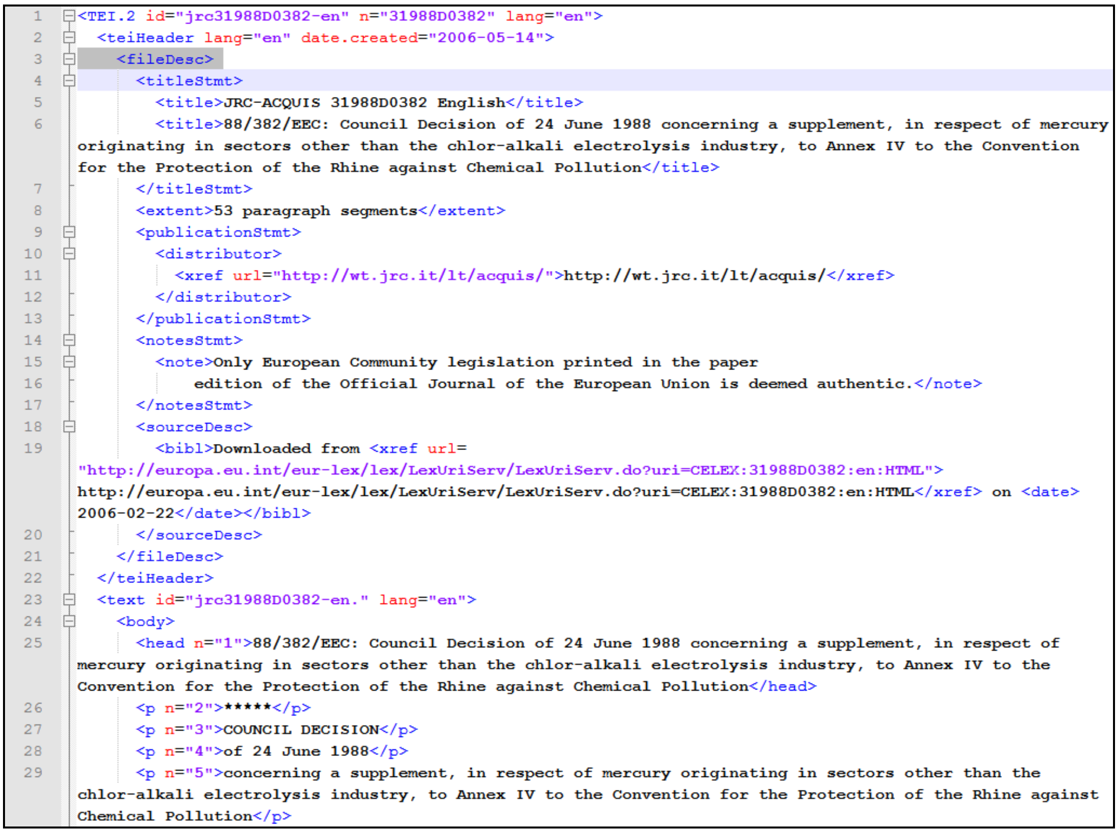
3.3.3.1 Pre-processing

Version 2.2 of JRC-Acquis consists of 15,532 XML files (an example file is shown in Figure 7) per language each of which contains an official legislation document of the European Union. Each of these files has its own version in each of the 20 languages of the EU. For each language pair there is also another XML file (part of the file is shown in figure 8) that connects lines for each pair of XML files. This file was used to align the corpus. Details of the JRC-Acquis corpus are shown in Table 3.

	Characters	Words	Paragraphs
English	41,917,357	6,465,374	227,007
Greek	46,013,152	6,649,694	

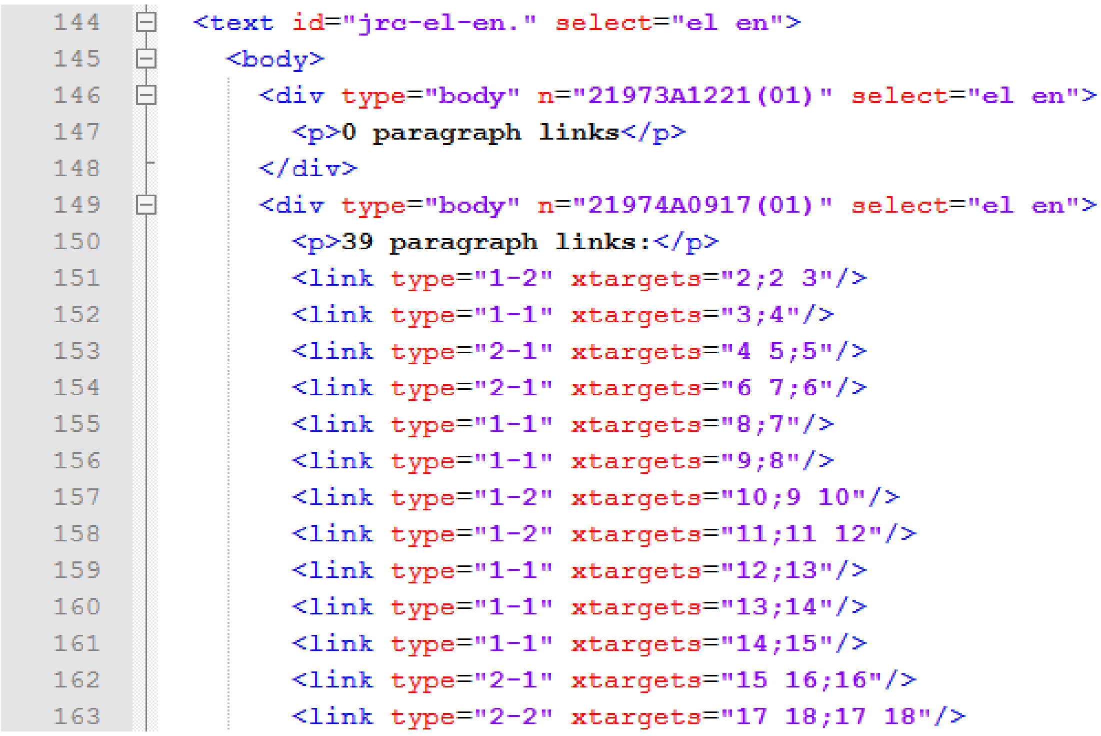
Table 3: The Greek-English JRC-Acquis corpus

JRC already had a perl script that transforms the XML files according to the alignment XML file. This script creates yet another XML file that contains both the source and target sentences. The input data required by MOSES are two files (one in the source and one in the target language) in plain text and paragraph aligned. In order to conform to these requirements, a program was written to transform all those XML files into two text files. It was written in Visual Basic and no XML parser class was used as it would pose a great overhead on the execution time of the program. The algorithm is shown below in Figure 9.



```
<TEI.2 id="jrc31988D0382-en" n="31988D0382" lang="en">
  <teiHeader lang="en" date.created="2006-05-14">
    <fileDesc>
      <titleStmt>
        <title>JRC-ACQUIS 31988D0382 English</title>
        <title>88/382/EEC: Council Decision of 24 June 1988 concerning a supplement, in respect of mercury
originating in sectors other than the chlor-alkali electrolysis industry, to Annex IV to the Convention
for the Protection of the Rhine against Chemical Pollution</title>
      </titleStmt>
      <extent>53 paragraph segments</extent>
      <publicationStmt>
        <distributor>
          <xref url="http://wt.jrc.it/lt/acquis/">http://wt.jrc.it/lt/acquis/</xref>
        </distributor>
      </publicationStmt>
      <notesStmt>
        <note>Only European Community legislation printed in the paper
            edition of the Official Journal of the European Union is deemed authentic.</note>
      </notesStmt>
      <sourceDesc>
        <bibl>Downloaded from <xref url=
"http://europa.eu.int/eur-lex/lex/LexUriServ/LexUriServ.do?uri=CELEX:31988D0382:en:HTML">
http://europa.eu.int/eur-lex/lex/LexUriServ/LexUriServ.do?uri=CELEX:31988D0382:en:HTML</xref> on <date>
2006-02-22</date></bibl>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text id="jrc31988D0382-en." lang="en">
    <body>
      <head n="1">88/382/EEC: Council Decision of 24 June 1988 concerning a supplement, in respect of
mercury originating in sectors other than the chlor-alkali electrolysis industry, to Annex IV to the
Convention for the Protection of the Rhine against Chemical Pollution</head>
      <p n="2">*****</p>
      <p n="3">COUNCIL DECISION</p>
      <p n="4">of 24 June 1988</p>
      <p n="5">concerning a supplement, in respect of mercury originating in sectors other than the
chlor-alkali electrolysis industry, to Annex IV to the Convention for the Protection of the Rhine against
Chemical Pollution</p>
```

Figure 7: Beginning of the XML file containing one official European Union piece of legislation.



```
<text id="jrc-el-en." select="el en">
  <body>
    <div type="body" n="21973A1221(01)" select="el en">
      <p>0 paragraph links</p>
    </div>
    <div type="body" n="21974A0917(01)" select="el en">
      <p>39 paragraph links:</p>
      <link type="1-2" xtargets="2;2 3"/>
      <link type="1-1" xtargets="3;4"/>
      <link type="2-1" xtargets="4 5;5"/>
      <link type="2-1" xtargets="6 7;6"/>
      <link type="1-1" xtargets="8;7"/>
      <link type="1-1" xtargets="9;8"/>
      <link type="1-2" xtargets="10;9 10"/>
      <link type="1-2" xtargets="11;11 12"/>
      <link type="1-1" xtargets="12;13"/>
      <link type="1-1" xtargets="13;14"/>
      <link type="1-1" xtargets="14;15"/>
      <link type="2-1" xtargets="15 16;16"/>
      <link type="2-2" xtargets="17 18;17 18"/>
```

Figure 8: Part of the file describing the alignments between the various JRC-Acquis files

3.3.3.2 Problems with the JRC-Acquis Corpus

The alignment provided contained lines in the source files that were not aligned to any line in the target files and vice-versa. Putting a file with empty lines into the training pipeline caused many problems, as most steps of Moses expect no empty lines. Training could not succeed until those lines were removed, along with their corresponding ones in the other language's file.

The alignment for the JRC-Acquis corpus was obtained using Vanilla, a program written by Pernilla Danielsson and Daniel Ridings, which implements the *Church and Gale / Dynamic Time Warping* algorithm. The source and documentation of the program are available at http://nl.ijs.si/telri/Vanilla/. Looking the two files' alignment, one can observe that it is not very accurate. For a large portion of some file pairs, most of the lines are unaligned. Nevertheless, a very good BLEU score was obtained on JRC-Acquis (better than Europarl), leading to the conclusion that better alignment could help achieve even better translations.

Encoding was another significant problem. JRC-Acquis was encoded in ISO-8859-7 for Greek and ISO-8859-1 for English. The final versions of the two aligned files needed to be transformed in UTF-8, as this encoding is used for Europarl. Since these two corpora are going to be merged, it was imperative that they used identical encoding. The problem was easily solved by changing the encoding when creating the JRC-Acquis corpus files.

3.3.3.3 *Test Data*

Three test files each containing 2,000 lines were extracted from the two original files and removed from them. Further details on the test files can be seen on Table 5.

Lines in the original	TEST_1	TEST_2	TEST_3
Start	34,699	139,462	219,523
End	36,698	141,461	221,522

Table 4: Test data taken from the JRC-Acquis corpus

3.3.4 Medical Corpus

The medical corpus consists of several abstracts and articles from the "RISC Factors" and "Paediatrics" Journals published in Greece. These are medical journals concerned with cardiological matters and the medical care of children, respectively. Alignment and pre-processing of the data were conducted –in most part– manually. The documents consisting the corpus were in Microsoft Word format (doc), Rich Text format (rtf) and Portable Document format (pdf). All these files had to be converted to plain text format (with different tools each time), correlated with one another and aligned (the hardest and most time consuming task).

The whole process took up a large amount of the dissertation’s time. It was a very tedious process but of utmost importance, as bad alignment could destroy the tuning and decoding process and yield bad translations.

The size of the corpus was very small, although the data used for the Greek language model consists of far more data than the parallel corpus itself. This was due to redundant Greek text from the “Peadiatrics” journal that didn’t correspond to any English text.

	Characters	**Words**	**Paragraphs**
English	1,541,349	228,342	2,378
Greek	1,727,386	241,328	
Greek LM	21,778,272	3,109,470	146,242

Table 5: The Medical corpus

Chapter 4

Implementation

4.1. Resources

Most of the experiments were conducted on the four clusters available at the University of Edinburgh (UoE), named “Townhill”, “Hermes”, “Lutzow” and “Lion”. Those clusters are also used by many researchers, students and teaching staff at the UoE. Normal computers are highly inefficient for such a memory and processor demanding task as training and tuning an SMT system on so many data. Normal computers do not have multiple processors and memory nodes. This means that for an experiment that normally takes five days on the cluster, it would take more than a month to run on an average desktop computer, due to lack of memory and parallelism capabilities.

It must be pointed out that one experiment run can take up to five days, but usually takes three to four. Nevertheless, some experiments took more than two weeks to finish as they were failing to finish for several reasons:

1. Mistakes in the configuration files. There were times where one little spelling error in a configuration file could set back an experiment for a few days.
2. Mistakes in the input files. Many hours were lost trying to discover empty lines in the JRC-Acquis corpus. Also, much valuable time was lost trying to correctly build sgml files for the “nist-bleu” evaluation.
3. Flaws in the training and tuning process of “Moses”. Moses is a research project with only a couple of years of development. Much progress has been made and is being made. Nevertheless, there are

bugs and imperfections that lead to incomprehensible error messages or no error messages at all when training but especially tuning is terminated abnormally. In those cases, an experiment needed to be continued from the last correctly ran step.

4. Cluster failures. The two clusters that were capable of running big training files fast enough ("Towhnhill" and "Hermes") had many problems. Some of them included:
    a. Keeping jobs in the queue for a large time interval (more than twelve hours). After staying in the queue for so long, the experiment needed to be killed off and restarted with the "-continue" switch.
    b. Putting jobs in the Error state despite the fact that there was nothing wrong with them. Most of the times there were free slots available and files were where they were supposed to be. These errors needed to be cleared if spotted on time (ie. when not sleeping). Otherwise, there would remain in the error state for a good amount of time stalling the whole experiment.
    c. Staying in the "t" state forever. State "t" occurs between the "q" (in the queue) and the "r" (running) state. Sometimes, a job would remain in this "zombie" state for inexplicably long amount of time (more than twelve hours) and had to be killed eventually.
    d. Downtimes. There were occasions when one of the clusters would crash unexpectedly. Recovery times were sometimes unacceptable (20 hours). This fact also caused congestions to the other clusters as people tried to continue their experiments on clusters still working.

All those reasons made experimentation very hard and time-consuming. At one point, it was apparent that not all experimentation planned would be feasible. A choice needed to be made. Only a limited number of experiments would be possible to run. The choices made, are reflected on the experimentation chapter.

4.2. Baseline

We trained on each of the three corpora individually without any additional TMs or LMs for any of them. All of them were run with LM order 5, maximum sentence length 60 and using lexicalized reordering. Results from those runs are shown in Table 6, 7 and 8 respectively for each corpus.

EUROPARL	Average	Test file		
		dev2006	devtest2006	test2007
BLUE score	**28.35**	28.29	27.94	28.82
Length		1.002	1.008	1.000

Table 6: BLEU scores for Europarl (baseline)

JRC-ACQUIS	Average	Test file		
		TEST_1	TEST_2	TEST_3
BLUE score	**29.66**	30.38	28.35	30.24
Length		1.000	0.978	0.968

Table 7: BLEU scores for JRC-Acquis corpus (baseline)

MED	Average	Test file	
		MED_TEST_1	TEST_MED_2
BLUE score	**20.22**	27.14	13.31
Length		1.034	1.170

Table 8: BLEU scores for Medical corpus (baseline)

Looking at these results, we can observe that in the medical file, TEST_MED_1 has a noticeably higher score than TEST_MED_2. The reason is that the SMT model was tuned on the first test file. Due to the small size of the corpus, sparse data infiltrate the corpus, and the logical after-effect is to have this kind of behaviour.

It is also noticeable that the JRC-Acquis corpus has the best performance over the other two corpora. Alignment – at least for the Greek-to-English pair – is not very good. Comparing alignment quality for the tow corpora, JRC-Acquis seems to be worse. It is, thus, evident that the better quality can only be explained on the standard kind of the language used. The size of JRC-Acquis is considerably smaller than that of Europarl (almost 3 times), which should favour Europarl.

4.3. Method A – Merging TMs and LMs

4.3.1. Description of Method

As described above, time and infrastructure restrictions made the strict choice of the methods used inevitable. The idea in this line of experiments is to train each corpus individually and then try to combine the models created in a way to increase translation quality for the in-domain test files. Here we consider the medical corpus to be the domain whose translations need improvement. Experimentation showed that adding extra information on the corpus, translation quality is increased.

4.3.2. Results

It has been shown that using an in-domain language model can increase translation quality by more than 2 BLEU points [P. Koehn and J. Shroeder, 2007]. We confirmed the increase in performance by using the Europarl corpus along with the medical one for the translation model but only the medical for the language model. We always tune on the medical test file. Table 9 shows the results in BLEU.

E = Europarl M = Medical	TM (Translation Model)		LM (Language Model)		Average	Test file	
	E	M	E	M		TEST_MED_1	TEST_MED_2
Europarl + Med	Yes	Yes	Yes	Yes	**27.61**	34.77	20.45
Europarl + Med	Yes	Yes	No	Yes	**26.65**	33.98	19.32
Med (baseline)	No	Yes	No	Yes	**20.22**	27.14	13.31

Table 9: BLUE scores (in grey) using the Medical corpus as the small in-domain corpus and Europarl as the huge out-of-domain corpus

It is apparent here that just by using a huge training corpus added to a very small one, improves translation quality by a considerable amount. The difference of more than 6 BLEU points could be biased by the fact that the testing and tuning files are too small – only 20 lines. It is also possible that the large increase in performance was due to the fact that there were many unknown words in the test and tuning files that could not be covered by only the medical corpus. Adding the Language Model of Europarl further improves results by almost 1 BLEU point.

An attempt was made to interpolate the two language models in order to bias the one (Medical) over the other (Europarl). Though, this attempt was not successful due to the restrictions referred to in chapter 4.1 – experiments failed to tune. The same happened with an attempt to bias one translation model (Medical) over the other (Europarl).

4.3.3. Configuration File Choice

Someone might wonder how high is the evaluation on Europarl using this setup (i.e. tuning on the Medical test file). Well, it does not perform well. The average of Europarl performance when tuning on the Medical corpus, is about 10 BLEU points. Of course, it is not binding in any way to use the weights and translation table of the above run. There is a way to choose which weight and configuration file to use. “moses.ini” contains the weights and paths to the phrase tables and language models. The phrase table and language model constructed for the setup we described above gives very good translations for medical sentences.

So, is there a way to tell whether a new sentence or document should be translated using the medical configuration and weights or the Europarl configuration and weights? Well, there are many ways.

One simple way is to compare the new sentence or document with the medical corpus and the Europarl corpus and see to which it resembles more. This can be easily done by running the “ngram” command of SRILM against each language model, check which perplexity is lower and choose the “moses.ini” file accordingly. So, whenever a medical document arrives, the settings for the best medical performance will be chosen, to get the best possible translation. The same will happen when a document that more resembles Europarl arrives to the input of our system.

Another way is using a simple machine learning algorithm – for instance Naïve Bayes – to classify the new document into the category of “Europarl” or “Medical”. There are many good classification algorithms and tools out there that can be adjusted for the particular task.

4.4. Method B – Clustering

4.4.1. Description of Method

Categorizing all the documents of a bilingual corpus in predefined classes was the first idea when starting the thesis. It was hard, though, to get a classification scheme that would have a good balance between coarse and fine-grained categories. Then, the Eurovoc classification was come across. Eurovoc[2] is a Thesaurus developed by the European Parliament and the EC’s Publications Office (OPOCE), together with several national organizations.

Its advantages include a fully controlled vocabulary with wide coverage. It is a multilingual thesaurus – exists in all 21 official languages of the EU plus

[2] http://europa.eu/eurovoc/sg/sga_doc/eurovoc_dif!SERVEUR/menu!prod!MENU?langue=En

Croatian, Russian, Albanian and Ucranian. It is organised hierarchically into a maximum of 8 levels. The top level contains 21 fields while the next one contains 127. The top-level categories of Eurovoc are shown in Figure 9.

04 POLITICS

08 INTERNATIONAL RELATION

10 EUROPEAN COMMUNITIES

12 LAW

16 ECONOMICS

20 TRADE

24 FINANCE

***28 SOCIAL* QUESTIONS**

32 EDUCATION AND COMMUNICATIONS

36 SCIENCE

40 BUSINESS AND COMPETITION

44 EMPLOYMENT AND WORKING CONDITIONS

48 TRANSPORT

52 ENVIRONMENT

56 AGRICULTURE, FORESTRY AND FISHERIES

60 AGRI-FOODSTUFFS

64 PRODUCTION, TECHNOLOGY AND RESEARCH

66 ENERGY

68 INDUSTRY

72 GEOGRAPHY

76 INTERNATIONAL ORGANISATIONS

Figure 9: The 21 categories of Eurovoc

Eurovoc also has an immediate connection to the JRC-Acquis corpus. Every European Union legislation document is indexed according to Eurovoc. Eurovoc is freele available for research purposes. Although a formal request was made at an early stage, the answer came only about a month later. The

Eurovoc XML document would take another two weeks to reach our hands. There was not enough time to use it properly. In the light of these developments, a new plan was devised. The JRC-Acquis would be split into clusters of similar documents according to a clustering algorithm.

The idea is to divide a big corpus into sub-corpora, then train and evaluate on each one of them. Then, when a new document needs to be translated, it is first classified into one of the predefined clusters and then using the translation and language model of the particular cluster to translate. [Yamamoto and Sumita, 2007]. This was also the original idea of the dissertation.

There are many issues that need to be addressed here. First of all, the method with which clustering is conducted. To simplify matters, a tree clusterer was chosen, CLUTO version 2.2.1. CLUTO uses simple algorithms to cluster anything that can be described with features. It is suitable for low and high dimensional datasets and tools to analyse the characteristics of the clusters created. The default clustering algorithm used is the Hierarchical Clustering Algorithm for Document Datasets [Zhao and Karypis, 2005].

4.4.2. Pre-processing

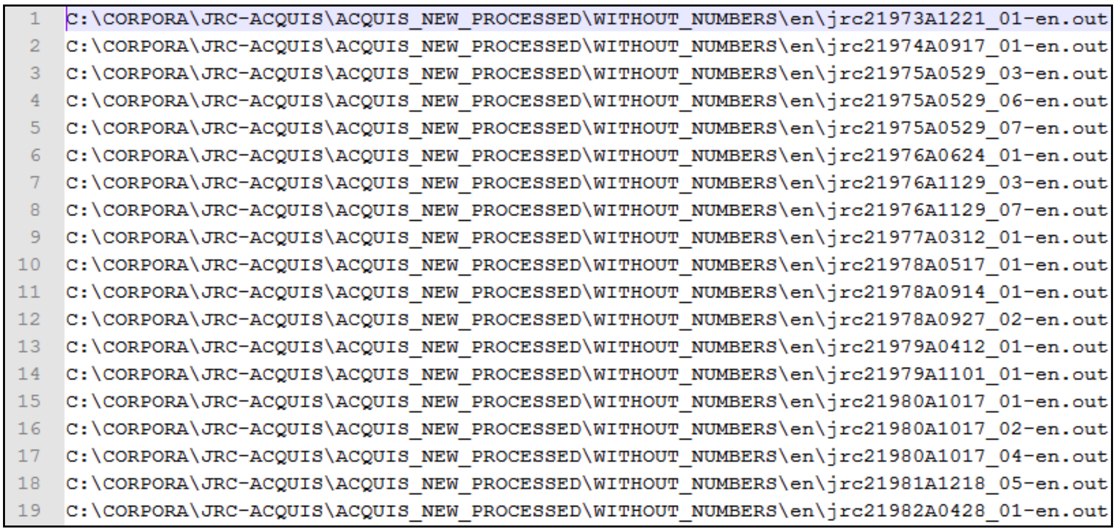
Some pre-processing was required in order to run CLUTO. The following procedure was used:

1. The Language Model of JRC-Acquis was taken.
2. All stop-words were removed from the LM.
3. Very common words in JRC-Acquis were removed (words like “article”, “number”, “EU”, “European”). These are words that occur to almost

every Acquis document – thus having a very high probability in the Language Model.

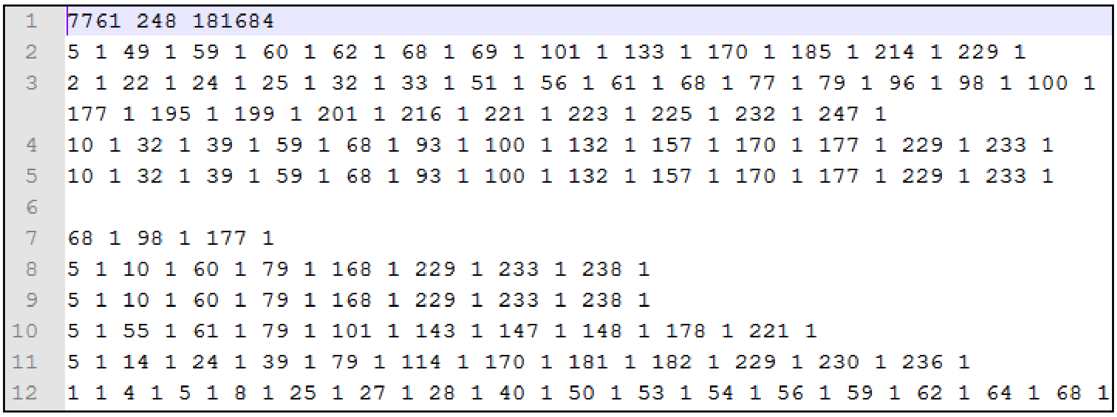
4. The first 248 of the remaining words were kept as the list to be used in clustering.
5. Two files were created using the program:
    a. One with 7761 lines containing the paths of the file names consisting the JRC-Acquis (one-dimensional). See part of the files in Figure 10.
    b. One with 7761 lines. Each line represents each file of JRC-Acquis. If one of the 248 words is present in the document, then the number of the word is written plus the value (in this case “1” for presence of the word). See part of the file in Figure 11.

The JRC-Acquis corpus was divided into 21 clusters. This choice of cluster numbers was based on EUROVOC thesaurus’ classification system. The run of CLUTO on JRC-Acquis is shown in Figure 12.



```
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21973A1221_01-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21974A0917_01-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21975A0529_03-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21975A0529_06-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21975A0529_07-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21976A0624_01-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21976A1129_03-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21976A1129_07-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21977A0312_01-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21978A0517_01-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21978A0914_01-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21978A0927_02-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21979A0412_01-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21979A1101_01-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21980A1017_01-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21980A1017_02-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21980A1017_04-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21981A1218_05-en.out
C:\CORPORA\JRC-ACQUIS\ACQUIS_NEW_PROCESSED\WITHOUT_NUMBERS\en\jrc21982A0428_01-en.out
```

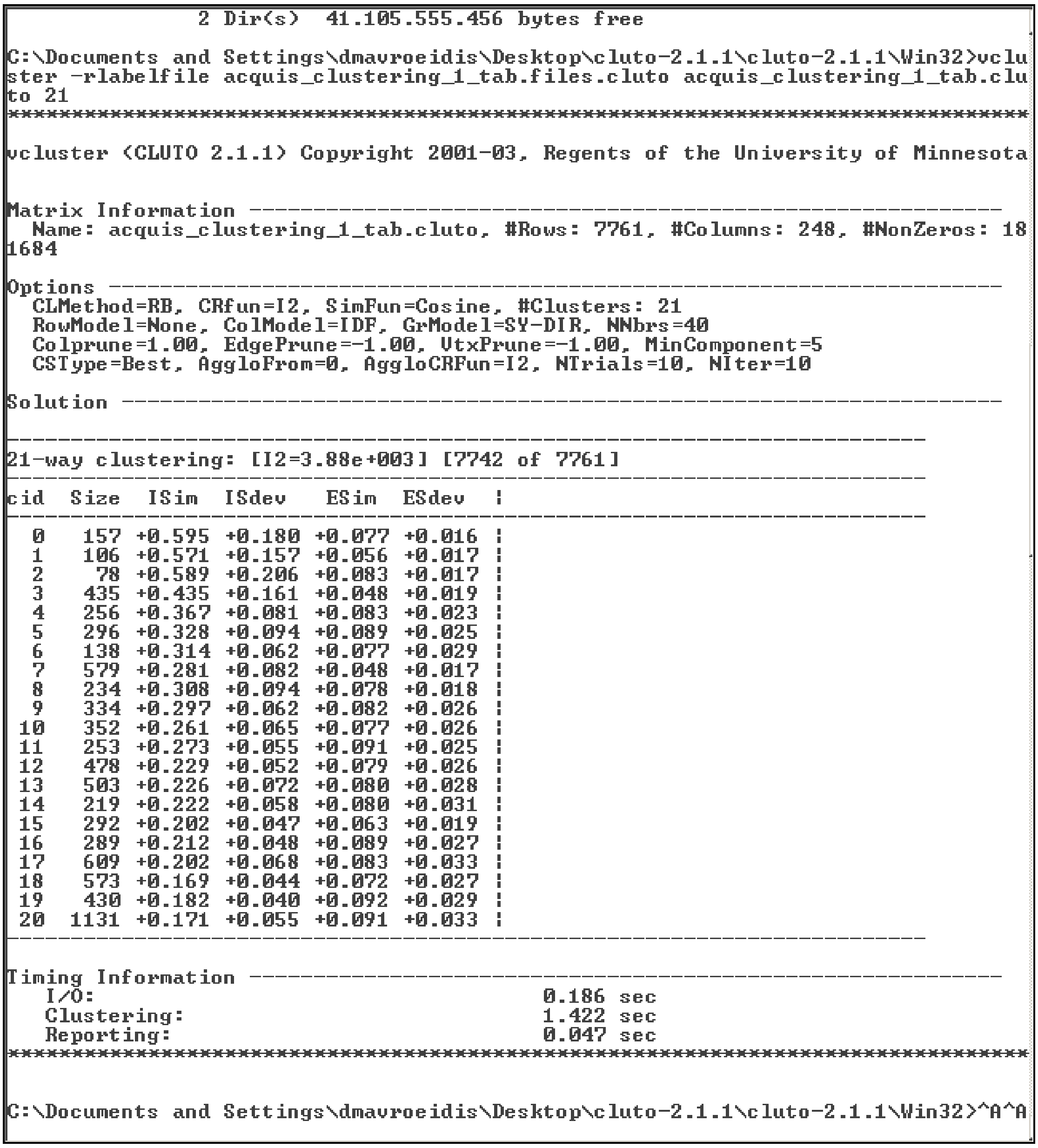
Figure 10: Part of the file containing all 7761 file paths of JRC-Acquis



```
7761 248 181684
5 1 49 1 59 1 60 1 62 1 68 1 69 1 101 1 133 1 170 1 185 1 214 1 229 1
2 1 22 1 24 1 25 1 32 1 33 1 51 1 56 1 61 1 68 1 77 1 79 1 96 1 98 1 100 1 177 1 195 1 199 1 201 1 216 1 221 1 223 1 225 1 232 1 247 1
10 1 32 1 39 1 59 1 68 1 93 1 100 1 132 1 157 1 170 1 177 1 229 1 233 1
10 1 32 1 39 1 59 1 68 1 93 1 100 1 132 1 157 1 170 1 177 1 229 1 233 1

68 1 98 1 177 1
5 1 10 1 60 1 79 1 168 1 229 1 233 1 238 1
5 1 10 1 60 1 79 1 168 1 229 1 233 1 238 1
5 1 55 1 61 1 79 1 101 1 143 1 147 1 148 1 178 1 221 1
5 1 14 1 24 1 39 1 79 1 114 1 170 1 181 1 182 1 229 1 230 1 236 1
1 1 4 1 5 1 8 1 25 1 27 1 28 1 40 1 50 1 53 1 54 1 56 1 59 1 62 1 64 1 68 1
```

Figure 11: Part of the file representation of JRC-Acquis files. The file descriptions correlate directly to the file paths in Figure 10



```
               2 Dir(s)  41.105.555.456 bytes free

C:\Documents and Settings\dmavroeidis\Desktop\cluto-2.1.1\cluto-2.1.1\Win32>vclu
ster -rlabelfile acquis_clustering_1_tab.files.cluto acquis_clustering_1_tab.clu
to 21
********************************************************************************

vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota


Matrix Information -----------------------------------------------------------
  Name: acquis_clustering_1_tab.cluto, #Rows: 7761, #Columns: 248, #NonZeros: 18
1684

Options ----------------------------------------------------------------------
  CLMethod=RB, CRfun=I2, SimFun=Cosine, #Clusters: 21
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10

Solution ---------------------------------------------------------------------

------------------------------------------------------------------------
21-way clustering: [I2=3.88e+003] [7742 of 7761]
------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  |
------------------------------------------------------------------------
  0   157 +0.595 +0.180 +0.077 +0.016 |
  1   106 +0.571 +0.157 +0.056 +0.017 |
  2    78 +0.589 +0.206 +0.083 +0.017 |
  3   435 +0.435 +0.161 +0.048 +0.019 |
  4   256 +0.367 +0.081 +0.083 +0.023 |
  5   296 +0.328 +0.094 +0.089 +0.025 |
  6   138 +0.314 +0.062 +0.077 +0.029 |
  7   579 +0.281 +0.082 +0.048 +0.017 |
  8   234 +0.308 +0.094 +0.078 +0.018 |
  9   334 +0.297 +0.062 +0.082 +0.026 |
 10   352 +0.261 +0.065 +0.077 +0.026 |
 11   253 +0.273 +0.055 +0.091 +0.025 |
 12   478 +0.229 +0.052 +0.079 +0.026 |
 13   503 +0.226 +0.072 +0.080 +0.028 |
 14   219 +0.222 +0.058 +0.080 +0.031 |
 15   292 +0.202 +0.047 +0.063 +0.019 |
 16   289 +0.212 +0.048 +0.089 +0.027 |
 17   609 +0.202 +0.068 +0.083 +0.033 |
 18   573 +0.169 +0.044 +0.072 +0.027 |
 19   430 +0.182 +0.040 +0.092 +0.029 |
 20  1131 +0.171 +0.055 +0.091 +0.033 |
------------------------------------------------------------------------

Timing Information -----------------------------------------------------------
   I/O:                                       0.186 sec
   Clustering:                                1.422 sec
   Reporting:                                 0.047 sec
********************************************************************************


C:\Documents and Settings\dmavroeidis\Desktop\cluto-2.1.1\cluto-2.1.1\Win32>^A^A
```

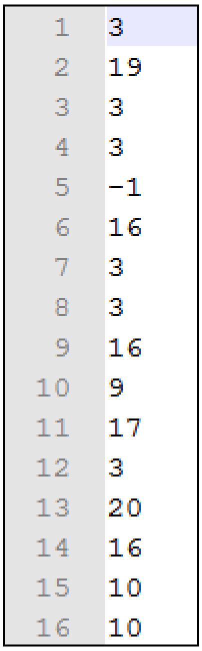
Figure 12: Running CLUTO on JRC-Acquis file descriptors

CLUTO uses the vector-space model [Salton, 1989] for representing documents. Each document is considered a vector in the word-space. Each document in our case is represented by the presence or not of a word in the document and not the frequency of its appearance. Similarity between documents is calculated using the cosine similarity metric which is defined as:

$$\cos(d_i, d_j) = d_i^t d_j / (\|d_i\| \|d_j\|) \qquad (4.1)$$

Clustering is conducted using the repeated cluster bisectioning approach [Steinbach et al., 2000]. When the algorithm starts executing, there are two clusters with documents. One of the two non-unary clusters are selected and divided into two new clusters. This process continues until the total number of clusters defined by the user is reached. There are several clustering criterion functions to perform the clustering. We use the default which maximizes the sum of the average pair-wise similarities between the documents assigned to each cluster weighted according to the size of each cluster [Zhao and Karypis, 2005] [Puzicha et al., 2000].

The output of CLUTO is shown in Figure 13. It just contains a list of numbers. These numbers denote the cluster in which each file belongs to.



```
3
19
3
3
-1
16
3
3
16
9
17
3
20
16
10
10
```

Figure 13: The output of CLUTO clustering process

By slightly altering the program of Acquis pre-processing, 21 new corpora were created. New test files were extracted, one for each cluster (200 lines each, from the beginning of each raw file – *head -200 file_name*). Training was conducted only on cluster_0 and cluster_18. Results were slightly as expected, though if trained on cluster_20 which was the largest cluster, results would be much better. First, the new corpus details are shown, but only for the two clusters used:

	Words		**Paragraphs**	
	cluster_0	**cluster_18**	**cluster_0**	**cluster_18**
English	94720	423453	3101	14874
Greek	99587	443414		

Table 10: The cluster_0 and cluster_18 corpus

Results from these runs are shown below:

C0 = cluster_0 A = JRC-Acquis	**TM (Translation Model)**		**LM (Language Model)**		**Test files**			
	A	**C0**	**A**	**C0**	**cluster_0_test**	**TEST_1**	**TEST_2**	**TEST_3**
cluster_0 + JRC-Acquis	Yes	Yes	Yes	Yes	**26.51**	19.85	19.57	23.11
cluster_0 + JRC-Acquis	Yes	No	Yes	Yes	**25.55**	25.98	25.28	27.41
cluster_0 + JRC-Acquis	Yes	Yes	No	Yes	**24.95**	14.01	10.29	13.65
cluster_0 + JRC-Acquis	Yes	No	No	Yes	**24.61**	14.06	11.29	14.78
cluster_0 (baseline)	No	Yes	No	Yes	**22.38**	14.36	9.99	13.27

Table 11: BLEU scores for cluster_0 sub-corpus

One can observe that replacing the translation model of cluster_0 with JRC-Acquis, we have a boost in BLEU score (+2). When adding the translation model for cluster_0 then have an additional increase in performance. This is very interesting, because cluster_0 is already inside the JRC-Acquis corpus.

This means that by adding the same corpus for the second time, we give it greater weight and translation quality is improved slightly.

Using both language models and only the JRC-Acquis training model, we get again a slight improvement.

Nevertheless, due to the small size of the cluster, performance never surpasses that of JRC-Acquis when ran as a whole.

C18 = cluster_18 A = JRC-Acquis	**TM (Translation Model)**		**LM (Language Model)**		**Test files**			
	A	**C18**	**A**	**C18**	**cluster_18_test**	**TEST_1**	**TEST_2**	**TEST_3**
cluster_18 + JRC-Acquis	Yes	No	Yes	Yes	**31.05**	28.07	27.29	29.12
cluster_18 + JRC-Acquis	Yes	Yes	No	Yes	**30.10**	25.14	19.98	24.06
cluster_18 + JRC-Acquis	Yes	No	No	Yes	**30.18**	24.73	21.57	24.72
cluster_18 (baseline)	No	Yes	No	Yes	**23.29**	21.60	16.55	20.01

Table 12: BLEU scores for cluster_18 sub-corpus

Chapter 5

Conclusions

More than 200 experiments were attempted with less than a quarter reaching a successful completion. Notwithstanding the computing difficulties, some very useful conclusions were drawn.

Training an SMT system on a combination of a small in-domain and a large out-of-domain corpus greatly improves translation quality when translating in-domain documents.

Additionally, translation quality can be improved by clustering a large corpus into smaller ones and building separate translation and language models. It is important, though, that the training should include the original corpus' translation and language model and tuned on a test file from the cluster. That way, if the cluster is big enough, it can surpass the original corpus in translation quality (BLEU score).

The number of clusters produced should be adjusted and balanced so that their size is big enough to yield good results but small enough to differentiate from the other clusters. In this dissertation, a constant number of clusters was taken because of time restrictions. It would be a good practice to experiment thoroughly as to what the best choice is before running experiments.

It would be interesting to have a machine learning classifier in order to put any new sentence in the correct cluster. Using the LM is a genial idea, but since the documents are already classified into clusters, training a classifier would be quite useful.

It seems like the most important component in training is tuning. When the model is tuned on test files of different genres, it produces very different results on the test files. Weights play the most significant role in the decoding process and that is where attention should be focused.

Last, but not least, JRC-Acquis - a fairly new multilingual corpus - was tried on Moses. That is a very important development, since the translation scores from JRC-Acquis are better than those of Europarl (by 1.3 BLEU score), despite the fact that JRC-Acquis is smaller in size.

Bibliography

Brown, P.F., Cocke J.,Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S., *A statistical approach to machine translation.* Computational Linguistics, 1990. **16**(2): p. 79-85.

Brown, P.F., Cocke J.,Della Pietra S. A., Della Pietra V. J., Jelinek F., Mercer R. L., Roossin P. S et al., *A statistical approach to language translation*, in *International Conference on Computational Linguistics (COLING)*. 1988.

Carpuat, M. and D. Wu. *Word Sense Disambiguation vs. Statistical Machine Translation*. in *43rd Annual Meeting of the ACL*. 2005.

Federico, M. and M. Cettolo. *Efficient Handling of N-gram Language Models for Statistical Machine Translation*. in *Second Workshop on Statistical Machine Translation*. 2007. Prague: Association for Computational Linguistics.

Jelinek, F., *Statistical methods for speech recognition*. Language, speech and communication. 1997, Cambridge, Mass., London,: MIT Press. xxi, 283 p.

Kay, M., *The Proper Place of Men and Machines in Language Translation*. 1980, Xerox Palo Alto Research Center.

Knight, K., *Decoding complexity in word-replacement translation models.* Comput. Linguist., 1999. **25**(4): p. 607-615.

Knight, K. and D. Marcu, *Machine translation in the year 2004*, in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*. 2004. p. 4.

Koehn, P., *Europarl: A Multilingual Corpus for Evaluation of Machine Translation.* 2002.

Koehn, P., *Statistical Machine Translation*. 2007.

Koehn, P., F.J. Och, and D. Marcu. *Statistical phrase-based translation*. in *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. 2003. Edmonton, Canada: Association for Computational Linguistics.

Lonsdale, D.W., A.M. Franz, and J.R.R. Leavitt. *Large-Scale Machine Translation: An Interlingua Approach*. in *Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*. 1994.

Nagao, M. *A framework of a mechanical translation between Japanese and English by analogy principle*. in *Artificial and human intelligence*. 1984. Amsterdam: North-Holland: A.Elithorn and R.Banerji.

Och, F.J., *Statistical Machine Translation: From Single-Word Models to Alignment Templates*, in *Informatiker*. 2002: Dienstag.

Och, F.J. and H. Ney, *A Systematic Comparison of Various Statistical Alignment Models.* Computational Linguistics, 2003. **29**(1): p. 19-51.

Papineni, K., S. Roukos, T. Ward and W-J Zhu, *BLEU: A method for automatic evaluation of machine translation*. in *40th Annual Meeting on Association for Computational Linguistics*. 2001. Philadelphia, Pennsylvania: Association for Computational Linguistics.

Paul, M., T. Doi, Y. Hwang, K Imamura, H Okuma, E. Sumita, *Nobody is Perfect: ATR's Hybrid Approach to Spoken Language Translation*. in *IWSLT*. 2005.

Puzicha, J., T. Hofmann, and J.M. Buhmann, *A theory of proximity based clustering: Structure detection by optimization*, in *PATREC: Pattern Recognition*. 2000, Pergamon Press. p. 617–634.

Salton, G., *Automatic text processing : the transformation, analysis, and retrieval of information by computer*. Repr. with corr. ed. Addison-Wesley series in computer science. 1989, Reading, Mass. ; Wokingham: Addison-Wesley. xiii, 530 p.

Shannon, C., *A mathematical theory of communication.* Bell System Technical Journal, 1948. **27**(3): p. 379-423.

Steinbach, M., G. Karypis, and V. Kumar, *A comparison of document clustering techniques*, in *KDD Workshop on Text Mining*. 2000.

Stolcke, A. *SRILM -- An Extensible Language Modeling Toolkit*. in *Intlernational Conference on Spoken Language Processing*. 2002. Denver.

Weaver, W., *Translation*, in *Machine Translation of Languages*, M. Press, Editor. 1949(1955), MIT Press.

Witten, I.H. and E. Frank, *Data Mining: Practical machine learning tools and techniques*. 2 ed, ed. M. Kaufmann. 2005, San Francisco.

Yamada, K. and K. Knight. *A syntax-based statistical translation model*. in *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. 2001. Toulouse, France.

Yamamoto, H. and E. Sumita. *Bilingual Cluster Based Models for Statistical Machine Translation*. in *oint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007. Prague: Association for Computational Linguistics.

Zens, R., F.J. Och, and H. Ney. *Phrase-based statistical Machine Translation*. in *Proceedings of the German Conference on Artificial Intelligence*. 2002.

Zhao, Y. and G. Karypis, *Hierarchical Clustering Algorithms for Document Datasets.* Data Mining and Knowledge Discovery, 2005. **10**: p. 141–168.