Stephen Armstrong & Colm Caffrey & Marian Flanagan (Dublin)

Translating DVD subtitles from English-German and English-Japanese using Example-Based Machine Translation

Contents

Abstract

Due to limited budgets and an ever-diminishing time-frame for the production of subtitles for movies released in cinema and DVD, there is a compelling case for a technology-based translation solution for subtitles (O'Hagan, 2003; Carroll, 2004; Gambier, 2005). In this paper we describe how an Example-Based Machine Translation (EBMT) approach to the translation of English DVD subtitles into German and Japanese can aid the subtitler. Our research focuses on an EBMT tool that produces fully automated translations, which in turn can be edited if required. We have chosen these language pairs as both are commercially significant.[1] We will seed the EBMT system with a corpus consisting of existing human translations from DVD to automatically produce high quality subtitles for audio-visual content. To our knowledge this is the first time that any EBMT approach has been used with DVD subtitle translation. Schäler et al. (2003: 88) propose that "the time is ripe for the transformation of EBMT into demonstrators, and eventually viable products". We attempt to answer their call with an EBMT approach to the translation of subtitles.

1 Introduction

One widely publicized example of poorly received subtitles is the translation of the Japanese subtitles of the film *Lord of the Rings: The Fellowship of the Ring*. Complaints about the quality of the subtitles from fans led to petitions being sent to the film's Japanese distributor and the director, Peter Jackson. The reason for this dip in translation quality was put down to time pressures imposed on the subtitler, as well as a lack of background knowledge needed to fully appreciate the film and its language (O'Hagan, 2003). This example clearly highlights how film subtitling is often dictated by the more lucrative market-driven component of the entertainment business. In this paper we firstly look at the background surrounding the need for research into the domain of DVD subtitle translation. Section 3 looks in detail at the reasons why we have introduced EBMT into this research, and also focuses on how EBMT compares to other translation technology. Section 4 describes the architecture of our EBMT

[1] Germany is traditionally a dubbing country unlike Japan, but DVD releases require subtitles in German.

system, including the make-up of our purpose-built corpus. Focusing on evaluation methods employed to test our system and current work-in-progress, Section 5 discusses our chosen evaluation method for this stage of research and presents system output, together with comments on the quality. Following on from this, Section 6 outlines future developments and the next stage of evaluation.

2 Motivation for Research

2.1 Research Background

Our research focuses on the feasibility of using EBMT to translate subtitles from English into a different target language. This research has come about due to our awareness of the pressures subtitlers are put under on a day-to-day basis, given the huge increase in DVD production since their introduction in 1997, and the work required to produce the required multilingual subtitle translations. Over one billion DVDs are produced per annum in the United States alone.[2] The demand on subtitlers to produce good-quality subtitles is at a record high, and carrying out research into how technology may assist the subtitler can only be advantageous to all concerned. Anecdotal evidence from the European and East-Asian markets suggests that subtitler rates, particularly for the DVD market, are continuing to drop across the board, driving experienced subtitlers out of the market while opening opportunities to those new to the profession; this has implications for quality.

Our aims are: to produce good quality DVD subtitles in German and Japanese translated from English; this will in turn assist the subtitler with the translation process, thus speeding up the process. The subtitles produced will be of a high standard, which means saving time and costs for the subtitling company, and relieving the pressures put on subtitlers to produce subtitles given the unreasonable time-frame and budget available.[3]

2.2 Languages

The two languages chosen for the first stage of research are German and Japanese. Both of these countries display extremely healthy economies in relation to DVD sales. In 2004, sales of DVDs in Germany grew by 63%, with 464 million units in total sold in Europe alone.[4] In 2003, DVD revenues exceeded cinema ticket sales in Japan.[5] These are also important languages in the field of subtitling. Japan is classified as a subtitling country, meaning all audiovisual releases will contain subtitles. In contrast, Germany is primarily a dubbing country. However, all DVD releases in Germany are required to have German subtitles. A further rationale for using these two languages is their dissimilarity, which therefore allows us to fully test the applicability of EBMT for translating subtitles, as well as the robustness and the scalability of the system. Both languages have been the focus of previous MT research

[2] http://www.interactual.com/news/IRMA.htm [Accessed March 2006]

[3] There were heated discussions on the unreasonable time-frame and budget available for DVD subtitling at the *International Conference on Audiovisual Translation*: *In So Many Words*, held at the University of London from 6–7 February 2004. These were echoed also in the Languages and the Media conference held in Berlin, 3–5 November, 2004.

[4] http://www.variety.com/index.asp?layout=cannes2005&content=story&articleid=VR1117923182 [Accessed March 2006]

[5] http://www.nec.co.jp/press/en/0407/2601.html [Accessed March 2006]

(Carl & Way, 2003), however, neither have previously been used in the combined area of EBMT (which is itself a relatively new research area of MT) and subtitle translation.

3 Why EBMT?

First off, it might be better to address the question: "why Machine Translation (MT)?" Subtitles can be said to inherit some of the traits of what we call in the MT world a 'controlled language'. Usually a controlled language is characterized by simplified grammar and style rules, and a simplified and controlled vocabulary set. Certain constraints are imposed on the subtitler, such as the number of characters allowed per line, which may result in the subtitler choosing a more simplistic syntactic structure, while still conveying the original meaning. In addition to this, subtitles can be seen as a kind of transcription of spoken dialog (sometimes complete with obvious interjections such as ums and ahs). Both these factors infer that we should know a good deal about what kind of text is to be expected in subtitling, and that can only be positive for any translation task, including MT, as the more linguistic knowledge we have about the source language, the better the translation should be.

3.1 RBMT vs. EBMT

Some research has previously been carried out using Rule-Based MT (RBMT) for the translation of both closed captions (Popowich *et al*., 2000) and subtitles (MUSA IST Project) to varying degrees of success. However, recent research and development in MT show there is widespread belief that rule-based systems will never be good enough to warrant serious consideration in the domain of a controlled language. This is mainly due to lack of robustness and lack of coverage. With regards to subtitling and the similarities it shares with a controlled language, simpler syntactic structures (canonical forms) are often preferred as they tend to make sentences shorter, and thus more easily and quickly understood. Punctuation also differs greatly, and the subtitler must follow a number of rules which are not necessarily the same in natural language use. Some of these rules include the addition of 'sequence dots' at the end of a line, indicating the sentence is incomplete, italics are used to indicate foreign words, and subtitles only typed in uppercase are usually used when transferring a display, such as a written signpost. RBMT assumes that the input sentence will be grammatically correct, however, subtitles will vary greatly from grammatically correct structures to sentences ridden with ellipses. They will also contain plenty of slang words which may not exist in the hand-coded RBMT dictionary. As EBMT relies on previously translated examples, it should be able to cope with both problems mentioned above.

3.2 TM vs. EBMT

Translation memory (TM) systems have become ubiquitous in making the translation process more efficient, and have been adopted by many of the big players in the localization industry. However, there is still an unwarranted tendency for Joe-Freelancer to be wary of said systems: it should be noted that these systems do not translate, they propose previously suggested human-translated 'examples' from a database, and it is up to the human to either accept or reject the suggested match. In other words, TM can basically be considered as a sophisticated search-and-replace engine (Schäler *et al*., 2003), and needs a human presence at all times during the translation process.

Drawing some parallels with TM but with some distinct differences is the notion of EBMT which goes back as far as the 1980s (Nagao, 1984). Here, like with TM systems, we

rely on a bilingual corpus, aligned at sentential level. In addition to this, EBMT goes a step further and goes beneath sentence level, 'chunking' each sentence pair and producing an alignment of sub-sentential chunks. Going beyond the sentence means we should have more scope for capturing useful matches which may be missed otherwise. EBMT is based on the principle of recombining these chunks to produce an automatic translation.

4 Our System

Our first step was to gather together a suitable corpus (described in section 4.1). We clean the data, split it into sentences, storing these sentences for later use. Our alignment program is run on these files, which results in a sententially-aligned bilingual corpus. The next step is to split sentences up into smaller units. For this we implement the Marker Hypothesis, which states that 'all natural languages are marked for complex syntactic structure at surface form by a closed set of specific lexemes and morphemes which appear in a limited set of grammatical contexts and which signal that context' (Green, 1979). Closed-class sets of words can predict or indicate what word classification will appear next. This is the basis for our system, and how we break up sentences and recombine them to generate new output, for example:

German subtitle:

EN: Did <PRON> you ever consult <PRON> that private detective?
DE: Waren <PRON> sie <ADV> eigentlich <PRON> bei <PRON> diesem Privatdetektiv ?

The resulting sub-sentential chunks would be the following, as chunks must contain at least one non-marker word. When a marker word is followed by another marker word, these two words are combined into one segment.

<PRON> that private detective? ⇔ <PRON> bei diesem Privatdetektiv?

These smaller units are stored in the corpus as aligned segments, so if an input sentence cannot be matched against a complete sentence stored in the parallel corpus, the input sentence is then broken up into smaller segments or chunks, and the system then checks if these input chunks are already stored in the corpus. If so, the corresponding segment is retrieved and recombined with other segments or words to produce suitable output in a different language.

We use a number of statistically and linguistically motivated methods to find the most probable matches and recombine them to produce a successful translation. We also use a modular architecture, which means it should be easy to adapt the system to new language pairs. All that has to be changed is the bilingual corpus, along with a new set of marker words for that particular language pair. Figure 1 is a diagram to explain what happens when an English sentence is entered into the EBMT system.

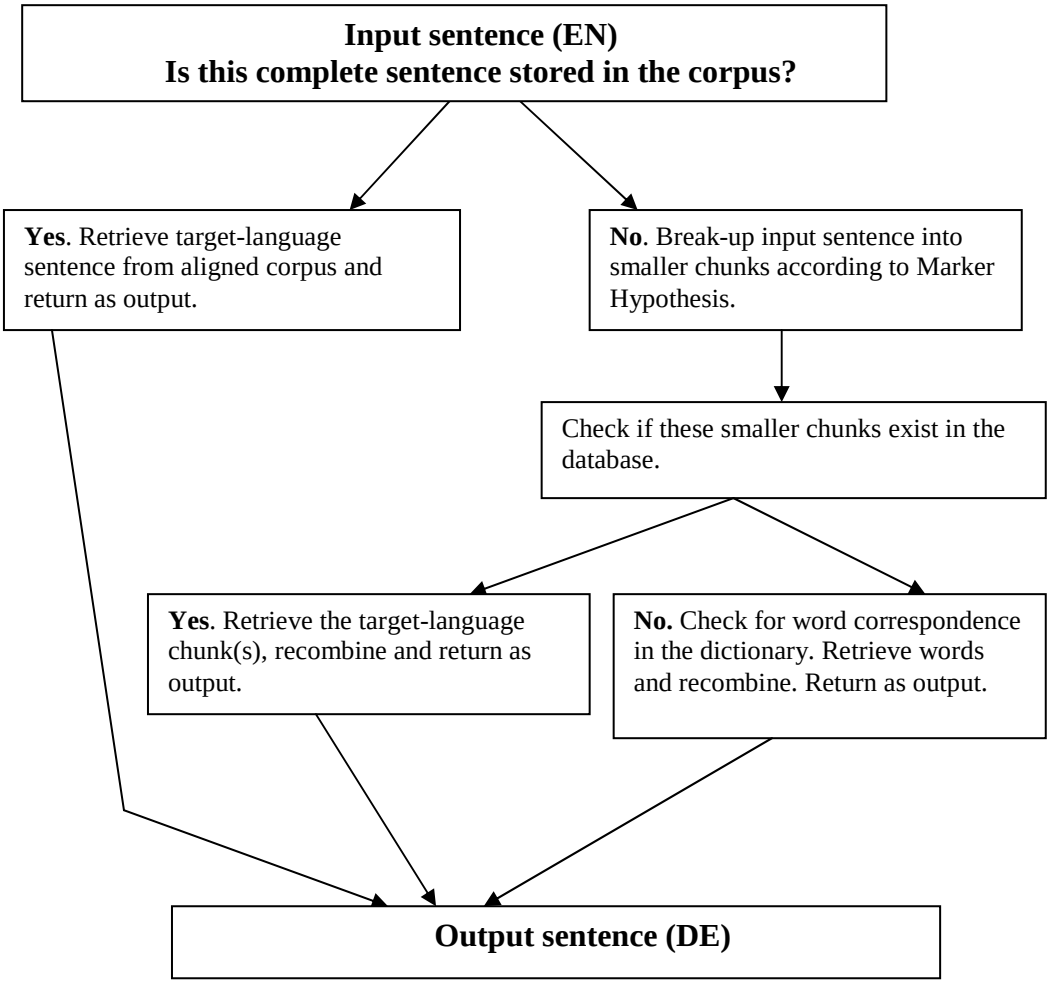

Fig. 1: EBMT system and the recognition of a sentence

4.1 Corpus

We concluded that the best way to do this was to build up a collection of DVDs which contain both English-German and English-Japanese subtitles. We extract these subtitles to text files using the freely available software SubRip, which gives us the subtitle text, along with their respective TC-in / TC-out (the time code at which the subtitle begins and ends).

So far we have extracted subtitles from almost 50 full-length features for the language pair English/German, which amounts to 64,996 sentences for English, and 61,292 sentences for German. Using the set of corpus linguistic tools, WordSmith, we were able to extract some interesting statistics from our own corpus. We calculated the average sentence length for both languages to be a little less than 6 words. Contrast this with the average length of sentences in for example the Europarl[6] corpus, which we calculated to be 24 words per sentence, and we can clearly see that our presuppositions about the language-constraints imposed on subtitler hold true. We plan to conduct a number of experiments on our corpus in an attempt to prove that the majority of language in the domain of film dialogs is in fact

[6] The Europarl corpus consists of thousands of sentences from proceedings in the European Parliament, and has been used in many EBMT research projects.

repetitive; from this we can deduce that there are plenty of reusable pre-translated chunks stored in memory that are available to the subtitler at the touch of a button.

Corpus collection for Japanese is moving along a little slower due to a number of factors. In Europe, DVDs with Japanese subtitles are a lot more difficult to come by as they are usually only sold in Japan (even Amazon.co.jp does not deliver these region-protected DVDs outside of Japan). Moreover, the OCR component of SubRip is not optimized to recognize Japanese characters. This means a lot of the work usually done automatically has to be done by hand, and takes hours rather than minutes to extract the subtitles for one movie. To date our English-Japanese corpus contains in excess of 10,000 sentence pairs, and will be analysed in a similar fashion to the English-German corpus.

4.2 System Requirements

In order to achieve these aims the system needs to meet particular criteria associated with any type of automated translation technology.

4.2.1 Acceptability

The system must produce subtitles, which are considered acceptable by the target-language audience. This acceptability level will be evaluated in the real-user evaluation pilot-study outlined in 6.1.

4.2.2 Readability

We will measure how readable the subtitles are by conducting some readability tests and also asking the opinion of participants of the pilot-study.

4.2.3 Grammaticality

The grammaticality of the target language is related to the acceptable level of subtitles. It is an important factor when evaluating the quality of the output. Given the fact the subtitles tend to be short, this will reduce the grammatical errors, which an MT system might have found difficult to deal with in the past. Short sentences are more likely to contain a less complicated grammatical structure than a longer sentence containing more than one clause. Thus this makes it much easier for a system to parse a short sentence correctly. It can also be the case that subtitles are not full sentences, leaving out the subject for example. This may be of benefit to the EBMT system, given the fact EBMT works on aligned sentences and segments of sentences previously translated by humans. These segments or ‘chunks’ could prove to be repetitive.

4.2.4 Efficiency

We aim to design a system that is efficient both in terms of speed and quality. The system is being used to translate DVD subtitles, so it is not real-time translation. That said, time is of essence when translating numerous DVDs, and therefore is still a high priority. Our EBMT system is programmed in Java, which is very efficient on memory.

5 Evaluation

5.1 Evaluation Methods

Real-user evaluation methods are the only reliable way of testing whether our system is working at a level, which would be suitable to produce subtitles for commercial markets, and accepted by the target audience. We are aware of previous evaluation studies involving web-based surveys by users of subtitles, for example the BBC 'Click Online' trials[7], but to our knowledge no real-user evaluations have been carried out before in the domain of EBMT automated subtitles, and within a home-entertainment setting.

User evaluation is essentially split into two types – *formative* and *summative*. Formative evaluation takes place during the development process, and is used to detect potential problems before the system is actually implemented (Preece, 1993). In contrast to this, summative evaluation is carried out when the system is finished (*ibid.*). This type of evaluation ensures the final system design is up to standard and is not used "to shape the design and development processes" (Landauer, 1997). In this paper we present formative evaluation results, which give us important input for improving the system. Other types of evaluation for our system include a pilot study within an audio-visual setting and the use of summative evaluation techniques, which will be carried out at a later stage of this study.

It is important to point out that the evaluation carried out at this stage of the research is preliminary work and at this stage no generalizations regarding the quality of our machine-translated subtitles can be made. These results aim to highlight what we need to do in order to improve the system. The evaluation presented here involved generating 2000 German sentences from our EBMT system. The Japanese system is not yet fully up and running, but will be included in our next stage of evaluation. From this German test set, we then randomly chose 200 sentences, and split these up into four groups of 50. The aim was to evaluate the intelligibility and accuracy of the automated German subtitles, by simply reading the MT produced subtitles printed on paper, and in no particular order. These automatic subtitles were from a selection of 30 DVD films. This type of formative evaluation, by simply reading the text and giving it a score is following Machine Translation evaluation protocol. It is a harsh way of evaluating subtitles, given the fact they are usually presented on a screen with the added influence of a picture and sound. However, it is also a very good method of highlighting areas we need to concentrate on in order to improve our system and the quality of the output. We would predict that the next stage of evaluation will benefit from this stage, with the introduction of audio-visual elements. It is often the case with a subtitled film that only a certain percentage of the understanding of a film is based on an understanding of the text alone (Gottlieb, forthcoming). There are times when a viewer may miss a subtitle for reasons relating to, for example, the image, reading speed or lack of understanding. However, the overall understanding of the film is not generally affected.

The intelligibility and accuracy scales were based on work by van Slype (1980: 7) and by Nagao as described in Jordan *et al*. (1993: 55), taken from Wagner (1998: 94). We approached intelligibility and accuracy in the following ways. According to Kenny (personal communication), "it should be possible to evaluate intelligibility without any reference whatsoever to the source text, so accuracy should not come into it"; a text can be completely intelligible but bear little resemblance to the source text. This sometimes happens in human translation, and we call it 'translation by invention'. She also made a point regarding measuring the accuracy of our output (*ibid*.) "accuracy, on the other hand, should be

[7] http://news.bbc.co.uk/1/hi/programmes/click_online/4421335.stm [Accessed August 2006]

ascertained independently from intelligibility." Tables 1 & 2 explain the scales and the range of scores possible.

	Intelligibility Scale*
1	gut verständlich
2	ziemlich verständlich
3	schwer verständlich
4	unverständlich

Tab. 1: Intelligibility Scale

	Accuracy Scale**
1	Satz übermittelt die Bedeutung des englischen Satzes
2	Satz übermittelt im großen und ganzen die Bedeutung des englischen Satzes
3	Satz übermittelt die Bedeutung des englischen Satzes nicht angemessen
4	Satz übermittelt die Bedeutung des englischen Satzes nicht

* & ** English translation of the German given in Appendix A

Tab. 2: Accuracy Scale

Therefore, when measuring the intelligibility of our automated output sentences, the participants were told to only refer to the German output, and when evaluating accuracy, they used both the original English source text (DVD intralingual subtitles) and the EBMT-generated German subtitles. We wanted German native speakers to carry out the evaluation, therefore ensuring that the EBMT subtitles were evaluated to a high standard. We emailed the evaluation sheet to 13 participants and received back 8 responses. All participants are above the age of 20, their mother tongue is German, and they have all completed a third-level education course. Table 3 shows the distribution of the sentence test sets among the 8 subjects. This shows that not all sentences were evaluated by all subjects, but it is possible to compare some responses between pairs of subjects for sets 1 – 3.

Test Set Number	**Number of Subjects**
1	3
2	2
3	2
4	1

Tab. 3: Distribution of test sets among the subjects

The idea behind this type of evaluation is to judge the quality of the subtitles purely based on the text. There are no audio-visual elements included, allowing the results from this type of evaluation to help feed into the system development.

5.2 Results

The sample size of our volunteers has led us to present the results of the evaluation in a qualitative and interpretive framework rather than quantitatively, showing positive aspects of EBMT subtitles, and where improvements are required. These focus on different areas, including creativity of the system within the recombination stage, as well as some errors, efficient areas of chunking in the system, along with mismatched chunks.

Table 4 presents an example where the system has given suitable output notably different to the human created subtitles.

English Original Subtitle	Shh, shh, shh! Alright children, now quiet.
German Original Subtitle	Okay, Kinder, nun seid ruhig.
German EBMT Subtitle	Scht, scht, scht! Gut Kinder, mehr Ruhe!

Tab. 4: Sentence 9, Set 1

We can see that the system has suitably translated the "shh" utterance as "scht" and instead of the adjectival form of "quiet", as chosen by the human subtitler, the system opted for the nominal form. Both German sentences make sense when read alone, although the EBMT sentence may benefit more from the contextualization that would be offered by the extra semiotic channels which would be present in an audiovisual evaluation. The sentence also provides proof towards the subjective nature of the evaluation process, as can be seen from Table 5. One plausible reason for Subject B's low scores could be the lack of context mentioned above.

	Intelligibility		Accuracy	
EBMT Output Subtitle	Subject A	Subject B	Subject A	Subject B
scht, scht, scht! gut kinder, mehr ruhe! (Set 1 Sentence 9)	1	3	1	3

Tab. 5: Evaluation scores for Sentence 9 Set 1

The system's creativity can also be seen in the following translation of the English subtitle "That's the last one", which in the human subtitle reads "Aber es war das letzte Mal" and the system translated as "Das ist das Letzte". In this instance, the EBMT output may seem like a more accurate translation, though it should be made clear that the lack of context also works against the original subtitles when seen in a vacuum as they are now.

An interesting example of the results we can get from the system is the translation of "I got the suitcase" (Sentence 40 Set 4), which in the original German subtitle is translated as "Ich habe den Koffer". The EBMT translation gives a more colloquial "Ich hab den Koffer", which could be seen as a more 'equivalent' translation of the colloquial "I got" we find in the

English subtitle. Table 6 shows examples of how the system correctly translates short subtitles, demonstrating the system's ability to segment sentences correctly and to reuse the chunks for different input.

English Original Subtitle	What's the matter, baby?
German Original Subtitle	Was ist los, Baby?
German EBMT Subtitle	Was ist los, Baby?
English Original Subtitle	I don't know
German Original Subtitle	Ich weiβ nicht.
German EBMT Subtitle	Ich weiβ nicht.
English Original Subtitle	I was grateful
German Original Subtitle	Ich war dankbar
German EBMT Subtitle	Ich war dankbar

*Tab. 6: short segments translated corre**ctly***

Given the fact we are in the first stages of evaluation, we are aware of some problems that exist with the chunking algorithm, the recombination stage and the dictionary generated during runtime. The evaluation has helped us address lexical errors in the system, such as missing words and mistranslations, both of which can be seen in Table 7 below. We see that some words, like "dig" and "wide", are apparently not being found in the dictionary, where "as" in mistranslated as "ace".

English Original Subtitle	to dig it twice as wide.
German Original Subtitle	Doppelt so groβ graben.
German EBMT Subtitle	An dig sie zweimal ace wide.

Tab. 7: Sentence 43, Set 2

Problems with the chunking phase of the system have also become evident, where a chunk from the English subtitle is mismatched with a chunk from the German sentence, as in the example given in Table 8.

English Original Subtitle	Now, Mr. Ewell, can you...
German Original Subtitle	Nun, Mr. Ewell, können sie...
German EBMT Subtitle	Nun, ich habe, können sie...

Tab. 8: sentence 44, Set 1

Here we see that "Mr. Ewell" has been falsely aligned with the chunk "ich habe", which has wrongly replaced Mr. Ewell in the EBMT output. This sentence scored a 3 for intelligibility (see Table 1) from both subjects and a 3 from one and a 4 from the other for accuracy (see Tab. 2).

6 Future Work

6.1 Pilot Study

Our next stage of evaluation will involve a pilot study into the acceptance of EBMT subtitles. A selection of short clips, subtitled in German, will be shown to native German speakers. The clips will consist of segments from English and Japanese language films, helping us ascertain whether source language knowledge has an effect on the acceptability levels of EBMT subtitles.

The subtitled clips will be shown to participants and they will be asked to evaluate the subtitles. Some clips will present a random mixture of EBMT subtitles and human output, while others will be subtitled entirely with EBMT output and others with only human output. The random mixture will allow us, but not the viewer, to know which subtitles are human produced and which are not. It will enable us to note any major differences in their acceptability. Questionnaires and retrospective interviews will be carried out with participants to garner their opinions on the subtitle quality and acceptability. A screen capture button will enable participants to take a freeze-frame still of any sections of a clip they find surprising or of poor quality and will be a useful reminder for the interview stage.

6.2 Corpus selection and Correlation with Automatic MT Metrics

We intend to carry out an investigation into the effects the corpus used to train our EBMT system has on the quality of the subtitles produced. To do this, our system will be trained on a variety of corpora, including subtitles only, bonus material only, a combination of the two, as well as the use of heterogeneous material (Armstrong, *et al.*, forthcoming). This will help to clarify whether it is more effective to train the system on a smaller amount of specific data, or a large amount of heterogeneous data not specific to subtitles. Within the EBMT community there is a divided opinion about the effectiveness of using either heterogeneous or homogeneous source material (cf. Denoual, 2005). The output will be evaluated using both human evaluators and MT metrics. One commonly used MT metric we will employ is BLEU (Papineni *et al*., 2002), which is based on the idea of measuring the translation closeness between a candidate translation and a set of reference translations with a numerical metric. A BLEU score is given between 0 and 1, with a score of 1 indicating a perfect match between the output translation and (parts of) the reference translation(s). MT experts treat the notion of a "gold standard" translation as normal practice, comparing the system output against this reference translation. The benefit of comparing both sets of results will show the real need for human evaluation of MT output in the audio-visual domain.

6.3 Showcase

We plan to showcase a working demo of our system at a conference in October of this year. This will allow people to see for themselves the quality of the EBMT subtitles and envisage ways in which the technology could be applied to the field of audio-visual translation.

7 Conclusion

In this paper we outlined the context in which the research direction was taken and our procedure for developing a tool to be used within the domain of subtitle translation. We also outlined how we intend to build a robust system incorporating the user evaluation results as an insight into how to improve the overall system, and thereby producing better quality subtitles. Our research is well motivated, given the current difficulties the subtitling industry is facing in relation to unrealistic time frames and decreasing budgets, leading to increased difficulties in attracting highly-trained staff. Our approach to this research is novel, as there are currently no available commercial EBMT systems. This could open up a new direction between audiovisual translation and technology.

8 References

Armstrong, Stephen & Caffrey, Colm & Flanagan, Marian. (forthcoming): 'Improving the Quality of Automated DVD Subtitles via Example-Based Machine Translation (EBMT)'. In *Translating and the Computer 28*. London: Aslib.

Carl, Michael. & Way, Andy (2003): *Recent Advances in Example-Based Machine Translation*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Carroll, Mary (2004): 'Subtitling: Changing Standards for New Media' [Online]. http://www.translationdirectory.com/article422.htm> [Accessed January 2006].

Denoual, Etienne (2005): 'The Influence of Example-Base Homogeneity on EBMT Quality'. In *Proceedings of the Second Workshop on Example-Based Machine Translation*. Phuket, Thailand: MT Summit X. 35-42.

Gambier, Yves (2005): 'Is audiovisual translation the future of translation studies? A keynote speech delivered at the Between Text and Image'. Updating Research in Screen Translation Conference. 27-29 October 2005.

Gottlieb, Henrik (2007): 'Multidimensional Translation: Semantics turned Semiotics'. Proceedings of the Marie Curie Euroconferences MuTra *'Challenges of Multidimensional Translation'* – Saarbrücken 2-6 May 2005.

Green, Thomas (1979): 'The Necessity of Syntax Markers. Two experiments with artificial languages'. *Journal of Verbal Learning and Behaviour* 18. 481-486.

Jordan, Pamela W. & Dorr, Bonnie J. & Benoit, John W. (1993): 'A first-pass approach forevaluating machine translation systems.' *Machine Translation* 8(1-2). 49-58.

Landauer, Thomas K. (1997): 'Behavioral Research Methods in Human-Computer Interaction'. In Helander, Martin & Landauer, Thomas K. & Prabhu, Prasad (eds): *Handbook of Human-Computer Interaction.* 2nd Edition. Amsterdam: Elsevier Science. 203-227.

MUSA IST Project [Online]. http://sifnos.ilsp.gr/musa/> [Accessed November 2005].

Nagao, Makoto (1984): 'A Framework of a Mechanical Translation between Japanese and English by Analogy Principle'. In Elithorn, Alick & Banerji, Ranan (eds): *Artificial and Human Intelligence.* Amsterdam. 173-180.

O'Hagan, Minako (2003): 'Can language technology respond to the subtitler's dilemma? - A preliminary study'. In *Translating and the Computer 25*. London: Aslib

Papineni, Kishore & Roukos, Salim & Ward, Todd & Zhu, Wei-Jing (2002): 'BLEU: a Method for Automatic Evaluation of Machine Translation'. In *Proceedings of the 40th acl*. 311-318. Philadelphia, PA.

Popowich, Fred & McFetridge, Paul & Turcato, Davide & Toole, Janine (2000): 'Machine Translation of Closed Captions'. In *Machine Translation* 15. 311-341.

Preece, Jenny (1993): 'A Guide to usability: human factors in computing'. Wokingham, UK - Massachusetts, USA: Addison-Welsey.

Schäler, Reinhard & Carl, Michael & Way, Andy (2003): 'Example-Based Machine Translation in a Controlled Environment'. In Carl, Michael & Way, Andy (eds): *Recent Advances in Example-Based Machine Translation*. Dordrecht, The Netherlands: Kluwer Academic Publishers. 83-114.

Slype, Georges van (1980): 'Bewertung des Verfahrens SYSTRAN für die maschinelle Sprachübersetzung bei der K.E.G.' In *Lebende Sprachen: Zeitschrift für Fremde Sprachen in Wissenschaft und Praxis* 25. 6-9.

Appendix A

	Intelligibility Scale*
1	Easily comprehensible
2	Comprehensible
3	Difficult to comprehend
4	Incomprehensible

	Accuracy Scale**
1	German sentence fully conveys the meaning of the English sentence
2	On the whole, the German sentence conveys the meaning of the English sentence
3	German sentences does not adequately convey the meaning of the English sentence
4	German sentence does not convey the meaning of the English sentence