

Translating DVD subtitles using Example-Based Machine Translation

**Stephen Armstrong, Colm Caffrey, Marian Flanagan, Dorothy Kenny,
Minako O'Hagan and Andy Way**

Centre for Translation and Textual Studies (CTTS),
School of Applied Languages and Intercultural Studies (SALIS)
National Centre for Language Technology (NCLT), School of Computing
Dublin City University

MuTra, Copenhagen May 2006

Outline

- Research Background
- Audiovisual Translation: Subtitling
- Computer-Aided Translation and the Subtitler
- What is Example-Based Machine Translation?
- Why EBMT with Subtitling?
 - Translation Memory vs. EBMT
- Our System
- Evaluation: Automatic Metrics and Real-User
- Experiments and Results
- Ongoing and future work

Research Background

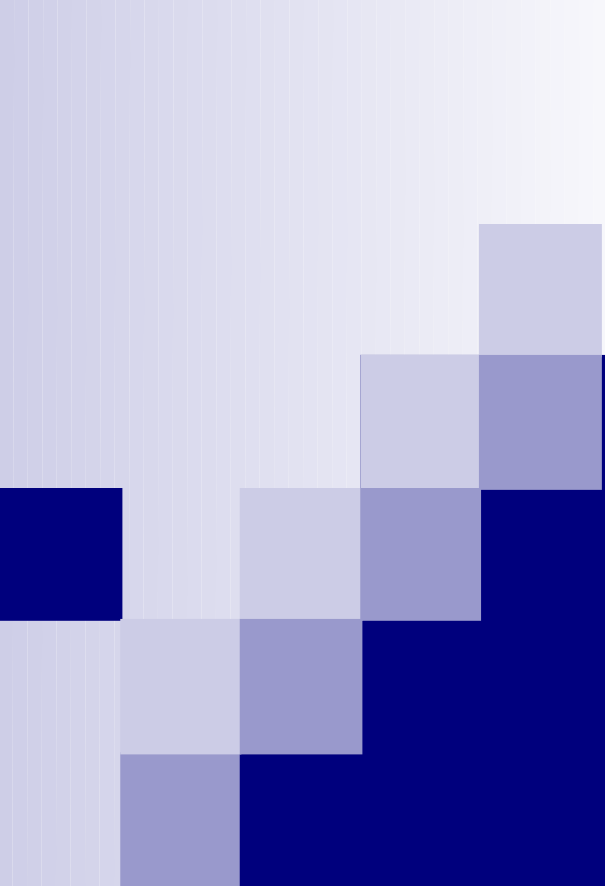
- One-year project funded by Enterprise Ireland
- Interdisciplinary approach

- Project idea developed from a preliminary study (O'Hagan, 2003)
- Test the feasibility of using Example-Based Machine Translation (EBMT) to translate subtitles from English to different languages
- Produce high quality DVD subtitles in both German and Japanese
- Develop a tool to automatically produce subtitles & assist subtitlers
- Why German and Japanese?
 - Germany and Japan both have healthy DVD sales
 - Dissimilarity of language structures to test our system's adaptability

- Recent research in the area
 - (O'Hagan, 2003) – preliminary study into subtitling & CAT
 - (Popowich et. al, 2000) – rule-based MT/Closed captions
 - (Nornes, 1999) – regarding Japanese subtitles
 - (MUSA IST Project) – Systran/generating subtitles

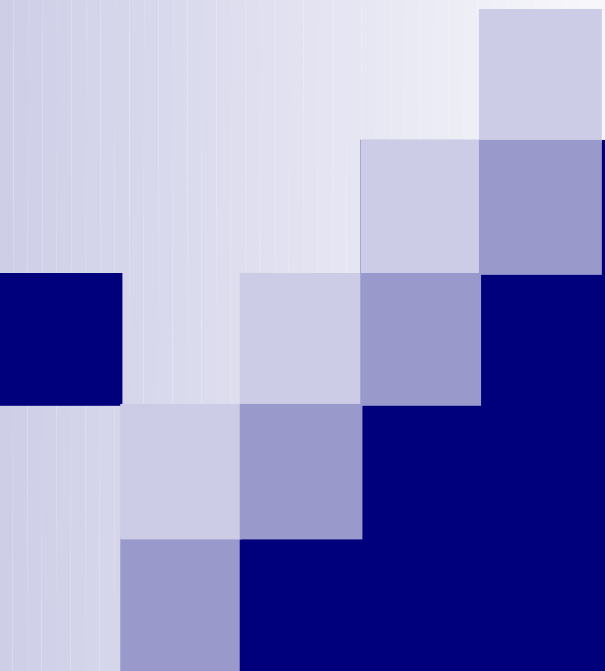
Audio-Visual Translation: DVD Subtitling

- As you are aware, subtitles help millions of viewers worldwide to access audiovisual material
- Subtitles are much more economical than dubbing
- Very effective way of communicating
- Introduction of DVDs in 1997
 - Increased storage capabilities
 - Up to 32 subtitling language streams
- In turn this has led to demands on subtitling companies



“The price wars are fierce, the time-to-market short and the fears of piracy rampant”

- (Carroll, 2004)

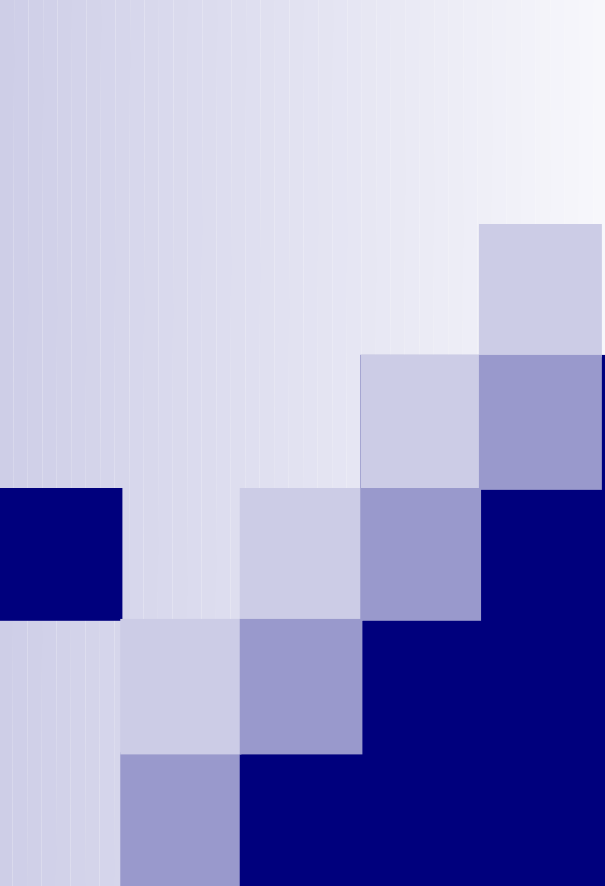


“One of the worst nightmares happened with one of the big titles for this summer season. I received five preliminary versions in the span of two weeks and the so-called 'final version' arrived hand-carried just one day before the Japan premiere.”

- Toda (cited in Betros, 2005)

Computer-Aided Translation (CAT) and the Subtitler

- Integration of language technology, e.g., Translation Memory into areas of translation like localisation.
- CAT tools have generally been accepted by the translating community.
- Proved to be a success in many commercial sectors
- However, CAT tools have not yet been used with subtitling software
- Some researchers have suggested that translation technology is the way forward



“Given limited budgets and an ever-diminishing time-frame for the production of subtitles for films released in cinemas and on DVDs, there is a compelling case for a technology-based translation solution for subtitles.”

- (O’Hagan, 2003)

What is Example-Based Machine Translation?

- Based on the intuition that humans make use of previously seen translation examples to translate unseen input
- It makes use of information extracted from sentimentally-aligned corpora
- Translation performed using database of examples extracted from corpora
- During translation, the input sentence is matched against the example database and corresponding target language examples are recombined to produce a final translation

Examples: EBMT

- Here are examples of aligned sentences, how they are “chunked” and then recombined to form a new sentence

Ich wohne in Dublin ⇔ I live in Dublin

Ich kaufe viele Sachen in Frankreich ⇔ I buy many things in France

Ich gehe gern spazieren mit meinem Ehemann ⇔ I like to go for a walk with my husband

Ich wohne in Frankreich mit meinem Ehemann ⇔ I live in France with my husband

Examples taken from (Somers, 2003)

The man ate a peach ⇔ hito ha momo o tabeta

The dog ate a peach ⇔ inu ha momo o tabeta

The man ate the dog ⇔ hito ha inu o tabeta

The man ate ⇔ hito ha ... o tabeta

the dog ⇔ inu

The man ate the dog ⇔ hito ha inu o tabeta

EBMT Example: Japanese

Input: She went to the tower to save us

Output: 彼女は私達を助けるために塔に行った
Kanojo ha Watashi-tachi wo Tasukeru-tameni Tou ni Itta

Source chunks:

今日彼女は買ったんだ (Sin City,
2005)

Kyō *Kanojo ha* *Katta-nda* ⇔ *She* bought it today

私達を狙ってる

Watashi-tachi wo Neratteru ⇔ He's after *us*

EBMT Example: Japanese (continued)

彼を助けるために君の才能を使え
2001)

(Moulin Rouge,

Kare wo **Tasukeru-tameni** Kimi no Sainō wo Tsukae ⇔ Use your talent **to save** him

塔の中で
2003)

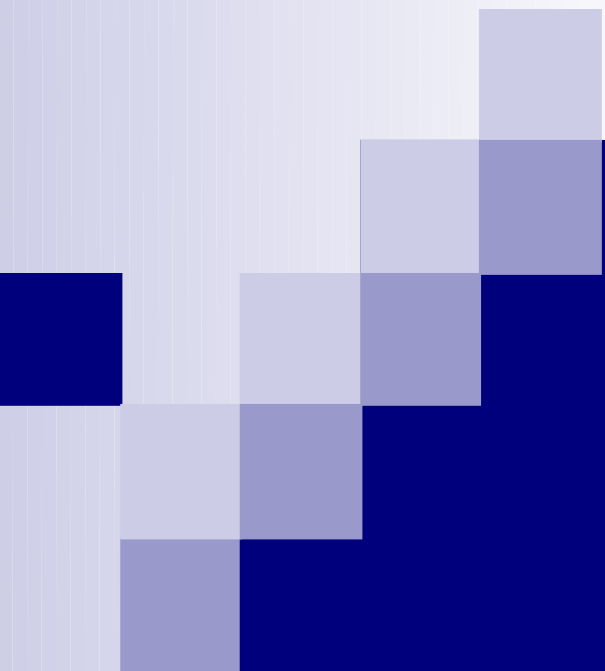
(Lord of the Rings,

Tou no Naka de ⇔ In **the tower**

君のアパートに行ったんだ
2005)

(Sin City,

Kimi no Apāto **ni Itta**-nda ⇔ We **went to** your apartment



“The Marker Hypothesis states that all natural languages have a closed set of specific words or morphemes which appear in a limited set of grammatical contexts and which signal that context.”

- (Green, 1979)

EBMT: Chunking Example

- Enables the use of basic syntactic marking for extraction of translation resources
- Source-target sentence pairs are tagged with their marker categories automatically in a pre-processing step:
- DE: Klicken Sie <PREP> auf <DET> den roten Knopf, <PREP> um <DET> die Wirkung <DET> der Auswahl <PREP> zu sehen
- EN: <PRON> You click <PREP> on <DET> the red button <PREP> to view <DET> the effect <PREP> of <DET>the selection

EBMT: Chunking Example

Aligned source-target chunks are created by segmenting the sentence based on these tags, along with word translation probability and cognate information:

- <PREP>auf den roten Knopf : <PREP> on the red button
- <PREP> zu sehen : <PREP> to view
- <DET> die Wirkung : <DET> the effect
- <DET> der Auswahl : <DET> the selection

- Chunks must contain at least one non-marker word - ensures chunks contain useful contextual information

Why EBMT with Subtitles?

- Based on translations already done by humans
- Subtitles also mainly used for dialogue
- Dialogue not always 'grammatical' so you need a **robust** system
- MT has been successful combined with controlled language
- Very few commercial EBMT systems
- Subtitles may share some traits of a controlled language
 - Restrictions on line length
 - The average line length in our DVD subtitle corpus is 6 words; comparing this with the Europarl corpus, which on average has 20 words per sentence
- However, in contrast to most controlled languages, vocabulary is unrestricted, necessitating a system with a wide coverage

Translation Memory (TM) vs. EBMT

- The localisation industry is translation memory-friendly, given the need to frequently update manuals
- Repetition is very evident in this type of translation
- Repetitiveness can be easily seen at sentence level
- Like TM, EBMT relies on a bilingual corpus aligned at sentence level
- Unlike TM, however, EBMT goes beneath sentence level, “chunking” each sentence pair and producing an alignment of sub-sentential chunks
- Going beyond sentence level implies ***increased coverage***

Evaluation: Automatic Metrics and Real-User

- Human evaluation will be used in conjunction with automatic metrics, such as BLEU
- Real-user evaluation of EBMT automated subtitles
- Subtitles generated by our system, then used to subtitle a section of a film on DVD
- Native-speakers of German and Japanese
- Real-user evaluation will consist of surveys, questionnaires and focus groups

Location

- Specially adapted translation research lab
- Wide-screen TVs pertaining to the setting of a cinema or home entertainment system

Experiments

- Different experiments involved different training & testing sets
 - DVD subtitles
 - DVD bonus material
 - Heterogeneous material (Europarl corpus, EU documents, News)
 - Heterogeneous material combined with DVD subtitles and bonus material
- To test which is the best corpus to use and which type of data to test the system
- Testing the system on a aligned corpus, German – English DVD subtitles, containing 11,000 lines
- Add in bonus material – 15,000 lines

Results

Subtitles taken from As Good As it Gets (1997)

- i need the cards (en)
- ich brauche die karten (de)
- ich brauche die karten (output)

- i'm sorry, sweetheart, but i can't (en)
- tut mir leid, liebbling, aber ich kann nicht (de)
- tut mir leid ,sweetheart, aber ich kann nicht (output)

- melvin , exactly where are we going (en)
- melvin , wo fahren wir denn hin (de)
- melvin , genau wo sind wir gehen (output)

Results (continued)

- you're welcome (en)
- gern geschehen (de)
- du bist willkommen (output)

- how is that a compliment for me (en)
- inwiefern ist das ein kompliment (de)
- wie ist das ein kompliment für mich (output)


- that 's just one idea (en)
- das war nur so eine idee (de)
- das ist nur eine idee (output)

Ongoing and Future work

- Continuous development of the EBMT system
- Continue building our corpus
- Investigate statistical evidence from our corpus
- Accurate description of the language used in subtitling
- Integration of system into a subtitling suite
- Automatic evaluation
- Real-user evaluation
- New language pairs
- Applications with minority languages
- Show proof of concept and moving on to the commercialisation phase

References

- **Betros, C.** (2005). The subtleties of subtitles [Online]. Available from: <<http://www.crisscross.com/jp/newsmaker/266>> [Accessed 22 April 2006].
- **Carroll, M.** (2004). Subtitling: Changing Standards for New Media [Online]. Available from: <<http://www.translationdirectory.com/article422.htm>> [Accessed January 2006].
- **Gambier, Y.** (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.
- **Green, T.** (1979). The Necessity of Syntax Markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behaviour* **18**:481-486.
- **MUSA IST Project** [Online]. Available from: <<http://sifnos.ilsp.gr/musa/>> [Accessed November 2005].
- **O'Hagan, M.** (2003). Can language technology respond to the subtitler's dilemma? - A preliminary study. *IN: Translating and the Computer* 25. London: Aslib
- **Nornes, A.M.** (1999). For an abusive subtitling. *Film Quarterly* **52** (3):17-33.
- **Fred Popowich, Paul McFetridge, Davide Turcato and Janine Toole.** (2000). Machine Translation of Closed Captions. *Machine Translation* **15**:311-341.



Thank you for your attention
Any questions? Feel free to ask

CTTS, SALIS

<http://www.dcu.ie/salis/research.shtml>

<http://www.ctts.dcu.ie/members.htm>

Dr Minako O'Hagan (minako.ohagan@dcu.ie)

Dr Dorothy Kenny (dorothy.kenny@dcu.ie)

Colm Caffrey (colm.caffrey@dcu.ie)

Marian Flanagan (marian.flanagan23@mail.dcu.ie)

NCLT, School of Computing

<http://www.computing.dcu.ie/research/nclt>

Dr Andy Way (away@computing.dcu.ie)

Stephen Armstrong (sarmstrong@computing.dcu.ie)