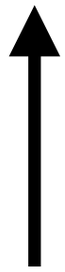
UiO : University of Oslo

Dialogue Modelling for *Statistical Machine Translation*

Pierre Lison
Language Technology Group

LT Seminar
28th January 2014

My current area of "expertise"



Dialogue Modelling for

Statistical Machine Translation

A new application domain

Many translation domains are conversational in nature:

… Subtitles for audiovisual content

Meeting transcripts

Speech-to-speech interpretation …

Many translation domains are conversational in nature

… but very little work has been done on machine translation (MT) specifically targeted for dialogue

↓

Project objective: use dialogue modelling to improve MT quality in conversational domains

Dialogue Modelling for

1. Demonstrate how to extract contextual features from the dialogue history…

Statistical Machine Translation

2. … and incorporate these features in the translation models of a MT system

Outline

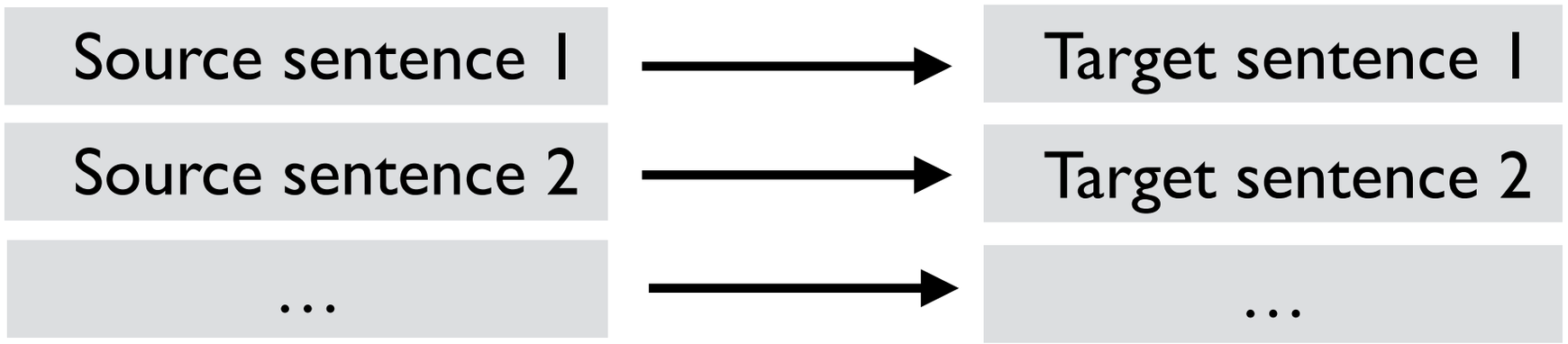
- **Motivation**
  - MT and the role of context
  - Context in dialogue translation
- **Proposed approach**
  - Source-side modelling
  - Target-side modelling
  - Implementation and evaluation
- **Practical aspects**

Outline

- Motivation
  - MT and the role of context
  - Context in dialogue translation
- Proposed approach
  - Source-side modelling
  - Target-side modelling
  - Implementation and evaluation
- Practical aspects

MT and the role of context

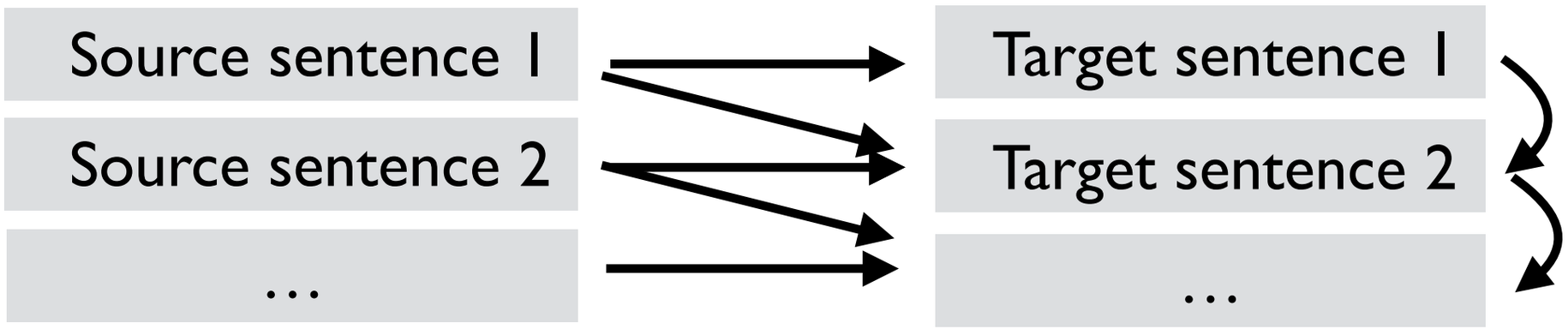
- Current MT systems translate sentences in isolation from one another
  - Source text viewed as unstructured bag of sentences
  - Easier for parameter estimation and decoding
  - But ignores the vast amount of linguistic information expressed at the cross-sentential level



MT and the role of context

- Renewed interest for discourse aspects of machine translation in recent years:
  - Contextual features in word-sense disambiguation
  - Discourse connectives
  - Lexical cohesion and consistency
  - Verbe tenses, pronominal anaphora

[see e.g. Hardmeier (2012) for a survey]



Context in dialogue translation

- Most research on discourse-oriented machine translation has focused on *text* materials (news articles, legal documents, etc.)
- Few have investigated how to exploit contextual factors in the translation of *conversational* material

That's where I enter the scene ;-)

Context in dialogue translation

- Key observation: dialogue is highly **cohesive**
  - Dialogue turns are tightly dependent on one another
  - What a speaker is saying at time *t* is often only interpretable in relation to the preceding history
  - Dialogue as a *collaborative activity*

→ Three examples to illustrate how dialogue context can affect the translation process

Example 1: Dialogue structure

A: Which way goes into town?
B: **Right.**

A: *Hvilken vei fører til byen?*
B: ***Høyre.***

A: So, those two don't work for Miletto. They work for Crenshaw.
B: **Right.**

A: *Så de to arbeider ikke for Miletto. De arbeider for Crenshaw.*
B: ***Riktig.***

[Source: OpenSubtitles parallel corpus]

Example 1: Dialogue structure

A: Which way goes into town?

B: **Right.**

Question > *Answer*

A: So, those two don't work for Miletto. They work for Crenshaw.

B: **Right.**

Statement > *Feedback*

Dialogue structure (dialogue act sequence) required to disambiguate "right"

Example 2: fragments

A: Mother… what was it like for you?
B: **For me?**

A: *Mor… hvordan var det for deg?*
B: ***For meg?***

A: You made this?
B: **For me?**

A: *Har du bygget den?*
B: ***Til meg?***

[Source: OpenSubtitles parallel corpus]

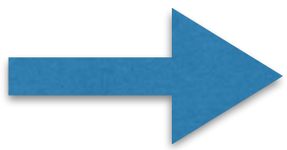
Example 2: fragments

A: Mother… what was it like for you?
B: **For me?**

experiencer

A: You made this?
B: **For me?**

beneficiary



Cross-sentential dependencies required to translate the preposition “for”

Example 3: Entrainment

A: Please, don't make the mistake of not taking me seriously, Roschmann.

B: I do **take you seriously.**

A: *Ikke gjør den feilen å ikke ta meg på alvor, Roschmann.*

B: *Jeg tar Dem på alvor.*

Reuse of expression "take X seriously"

→ **Dialogue history** required to find the most "salient" expression in the context

Outline

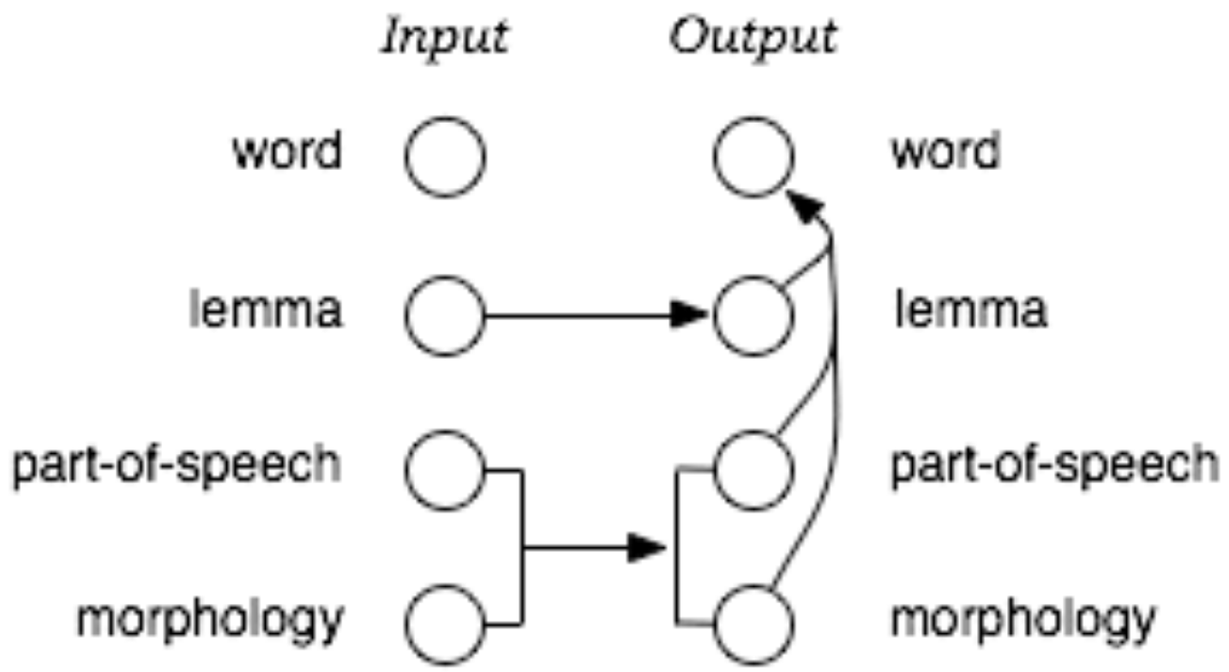
- Motivation
  - MT and the role of context
  - Context in dialogue translation
- **Proposed approach**
  - **Source-side modelling**
  - **Target-side modelling**
  - **Implementation and evaluation**
- Practical aspects

Proposed approach

- How can we exploit these insights to build better translation models?
- Work on both the source- and target-side of the translation process:
  - **Source-side**: extract new contextual features and integrate them in *(factored) translation models*
  - **Target-side**: strengthen the cohesiveness of the translations through *dynamic model adaptation*

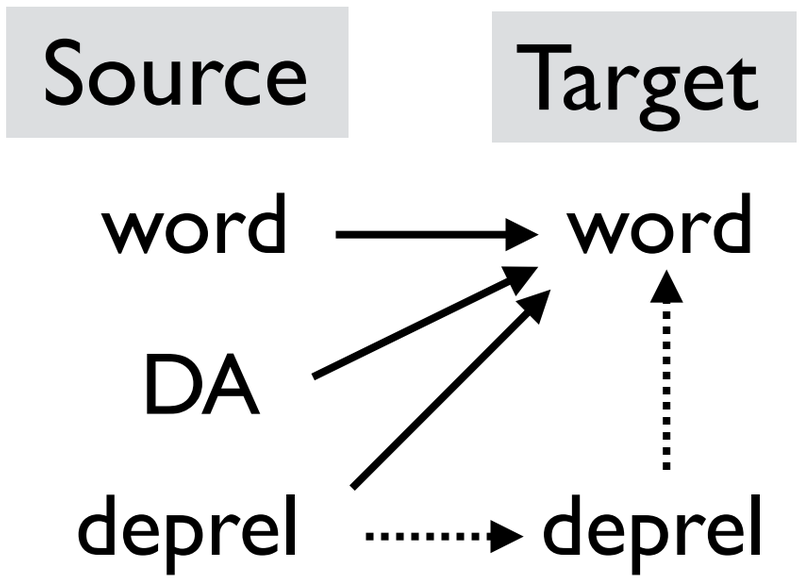
Source-side context

- Reliance on **factored** translation models
  - Extension of classical phrase-based models
  - Multiple layers of annotation for every word (token, lemma, POS, morphology, etc.)



Source-side context

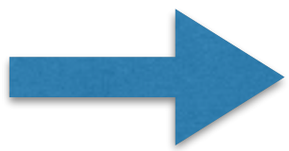
- We can use factored models to include additional dialogue features:
  - **Dialogue structure**: add a factor expressing the current dialogue act as predicted by a classifier
  - **Dependency relations**: add a factor expressing the (semantic) relation between the word and its head.



Source-side context: fragments

- Dependency relations would be especially useful to handle fragments (e.g. "for me?")
  - We first need to "reconstruct" the fragments (*Open question*: how to actually do that?)
  - And use a parser (and semantic role labeller?) to extract dependency relations from them

A: You made this?
B: **For me?**



A: You made this?
B: [**You made this**] **for me?**

Target-side context

- We can also use *target-side context* to ensure that the translations are coherent and consistent with one another.
  - Conversations are **cohesive**: only a few topics are "active" at a given time
  - **Grounding feedbacks** often reuse already uttered constructions (e.g. "I'm hungry" > "hungry?")
  - **Entrainment** also tend to increase the likelihood of already uttered constructions

Target-side context

- Dynamic model adaptation (in particular **caching techniques**) to strengthen the cohesiveness of the translations
  - Mix a static model (e.g. N-grams) with a dynamic model estimated from recent items in the history

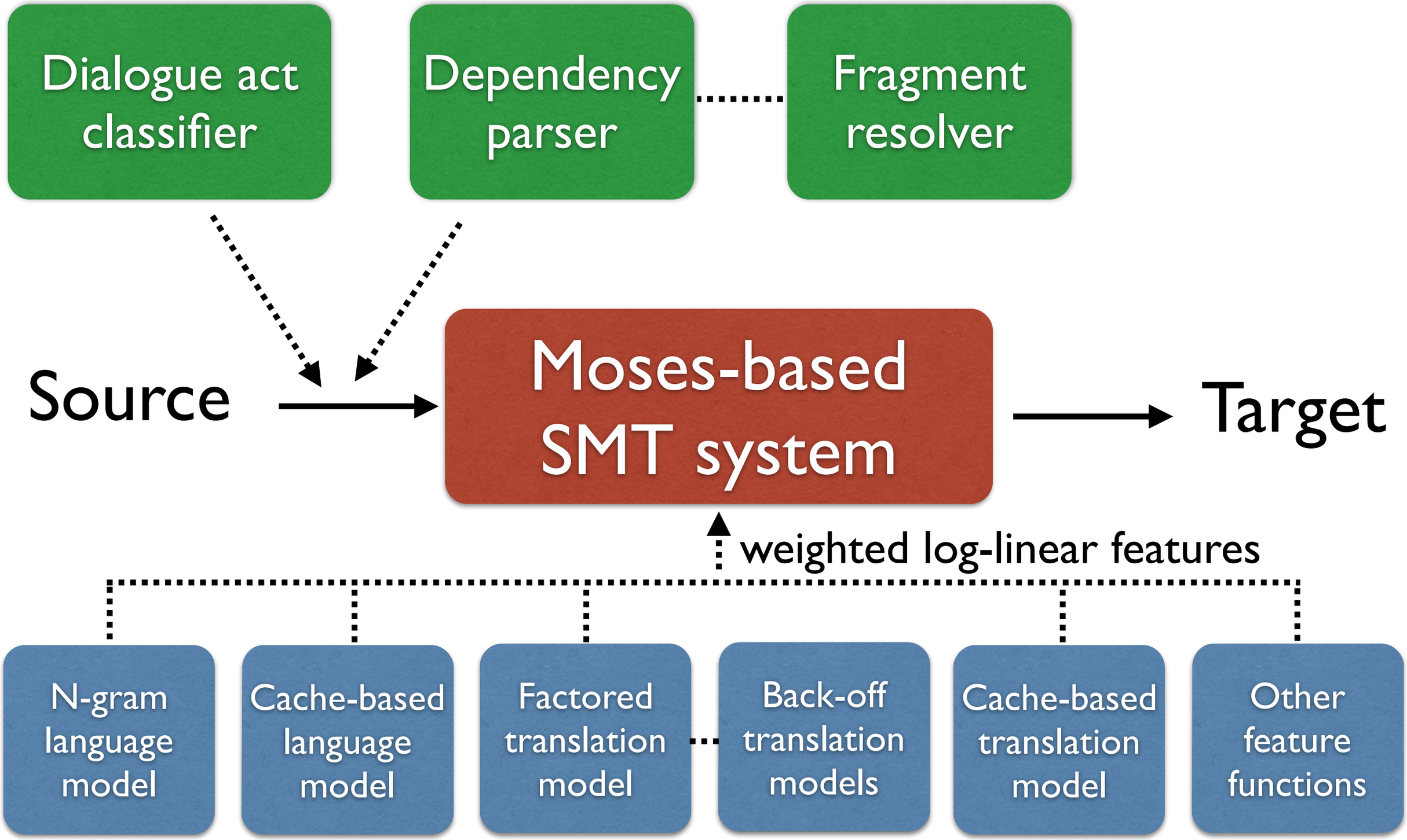
$$P(w_n|history) = (1-\lambda)P_{n\text{-}gram}(w_n|history) + \lambda P_{cache}(w_n|history)$$

- Can be applied for language and translation models

Target-side context

- Target-side context is harder to leverage than the source-side
  - The reference translations are not observable!
  - Risk of error propagation: erroneous translation for sentence *n* can affect downstream translations *n*+*1*, …
- Requires more advanced, document-level optimization techniques
  - Decoding in multiple passes, re-ranking

Implementation



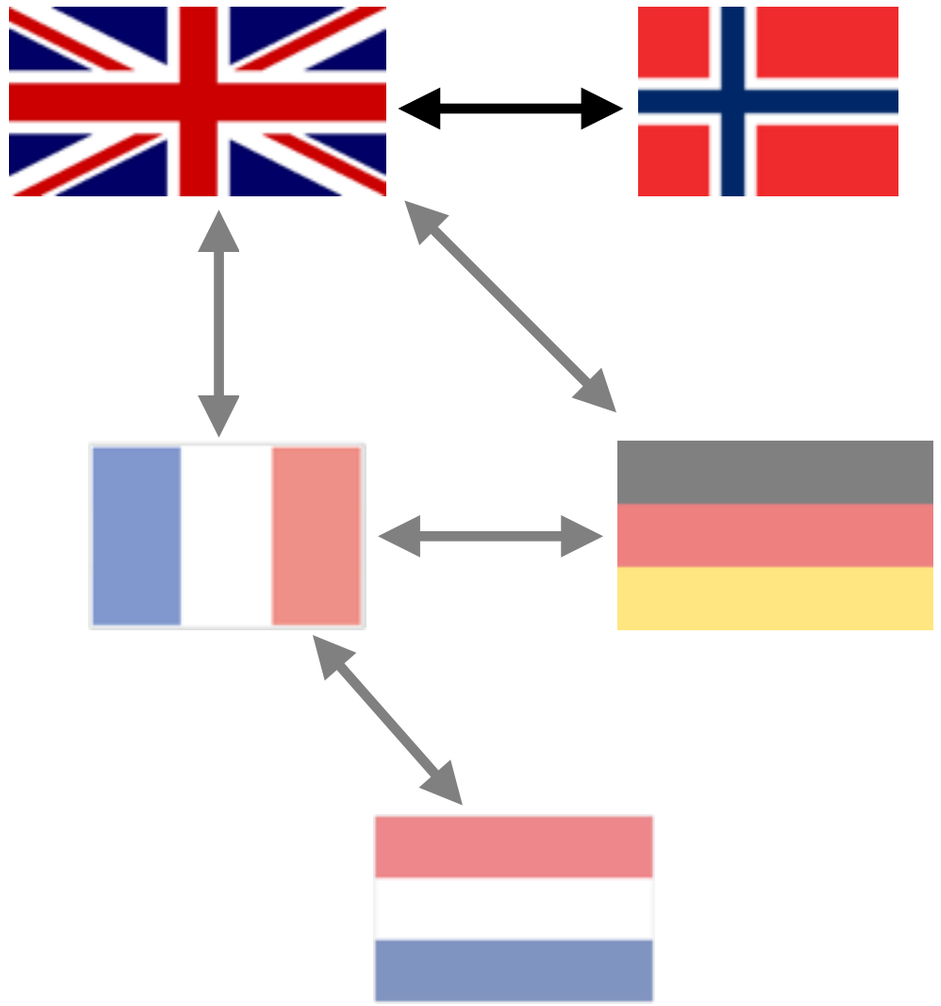
Evaluation

- Concrete application domain: **translation of subtitles**

  - Available training data
  - Real-world problem!
- Resources:
  - *OpenSubtitles* (fan-made, part of OPUS)
  - Repositories from *NRK* and *Broadcast Text*

Evaluation

- Language pairs:
  - English-Norwegian (if sufficient training data)
  - Other pairs (using French, Dutch, German, etc.)



- Evaluation metrics
  - Reference-based metrics such as BLEU and METEOR
  - More targeted evaluations for subcomponents

Outline

- Motivation
  - MT and the role of context
  - Context in dialogue translation
- Proposed approach
  - Source-side modelling
  - Target-side modelling
  - Implementation and evaluation
- **Practical aspects**

Practical aspects

- Project duration: **3 years**
- Funding from the *Norwegian Research Council*
  - FRIPRO funding scheme
- Bulk of the research conducted at LTG

Research stays

- Two planned research stays:
  - 3 months at **IDIAP** (Switzerland) to work with *Andrei Popescu-Belis* on source-side context models
  - 3 months in **Uppsala** to work with *Jörg Tiedemann* on target-side context models

Project partners

- Two project partners:
  - **NRK**
  - **Broadcast Text International** (subtitling & dubbing company)

- Will provide me with a privileged access to their archives of professionally translated subtitles

Project planning

Title	2014	2015	2016
• Setup resources and baseline system Preparation of resources,setup and evaluation of baseline system			
• Research stay at IDIAP, Switzerland (Spring 2014)			
Source-side modelling			
• Dialogue structure Integration of the dialogue act sequence in factored translation models			
• Utterance fragments Development of a new method to resolve utterance fragments, extract their dependency structure and integrate it in factored translation models			
• Preparation of long article on source-side modelling			
◆ Organisation of advanced PhD-level course on dialogue modelling			
◆ Intermediary milestone (first public release)			
• Research stay at Uppsala University (Autumn 2015)			
Target-side modelling			
• Caching techniques Construction of new cache-based models to account for dialogue cohesion			
• Global optimisation Work on decoding algorithms to optimise translations at the dialogue level			
• Preparation of long article on target-side modelling			
◆ Co-organisation of workshop on dialogue-oriented MT			
• Wrap-up and ultimate dissemination of results (concluding article)			
◆ Final milestone (second public release)			

Looking for collaborations!

Multiple "points of contact" with areas of current research @ LTG:

- Parsing "non-canonical" language
- Use of dependency features in statistical models
- Realisation ranking
- High-performance computing