Engaging Content
Engaging People

Automatic Construction of Discourse Corpora for Dialogue Translation

Longyue Wang
ADAPT Centre, Dublin City University
lwang@computing.dcu.ie

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Andy Way, Qun Liu

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund
SFI Science Foundation Ireland

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Outline

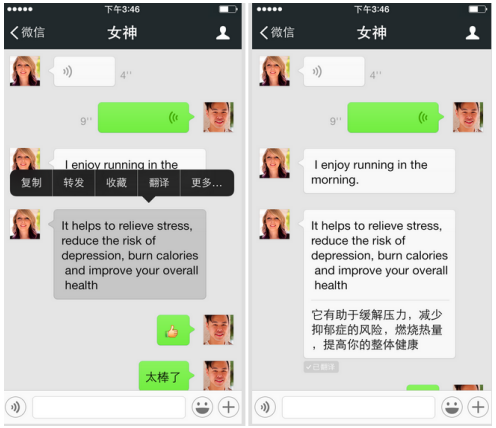
- **Motivation**
- **Related Work**
- **Methodology**
  - Examples
  - Proposed Approach
  - Results and Evaluation
- **Machine Translation Experiment**
  - Personalized dialogue SMT system
  - Results and Evaluation
- **Conclusion and Future Work**

Dialogue Machine Translation

Dialogue is an essential component of social behaviour to express human emotions, moods, attitudes and personality. **Machine translation** (MT) of conversational material products various real-life applications.

YouTube

WeChat



skype

Baidu 百度

Dialogue Machine Translation

We start a project on dialogue MT:

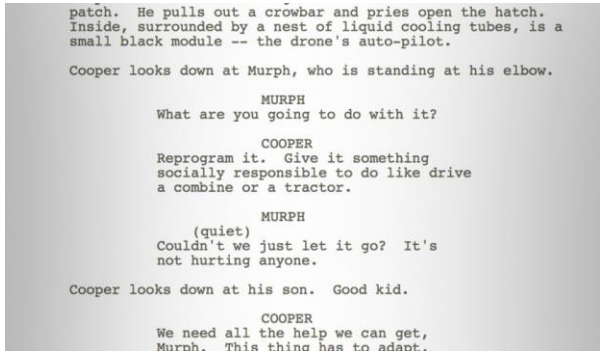
- Dialogue exhibits **more cohesiveness** than single sentence. Besides, it contains rich information such as specific structure, intention (dialog act, focus), speaker, subjective content (sentiment, agreement, decision, negotiation).
- To date, few researchers have investigated how to improve the dialogue MT by exploiting their **internal structure** or **collaborative activity**.
- Although there are a number of work on corpus construction for various natural language processing tasks, dialogue corpora are still scarce for MT.

Therefore, we propose a simple but effective method to automatically build corpora with rich information for exploring dialogue machine translation tasks.

Related Work

- Movie subtitles and scripts are commonly used for NLP tasks.
- Some work regard **bilingual subtitles as parallel corpora**, but it only focuses on single sentence (Tiedemann, 2012; Zhang et al., 2014). E.g., Lison and Tiedemann (2016) release OpenSubtitles2016.
- Other work focus on **internal structure** of dialogue from **movie scripts**. But these are monolingual data which cannot be used for MT (Walker et al., 2012; Schmitt et al., 2012). E.g., Hu et al. (2013) release Internet Movie Script Database (IMSDb).

Movie Subtitles


patch. He pulls out a crowbar and pries open the hatch. Inside, surrounded by a nest of liquid cooling tubes, is a small black module -- the drone's auto-pilot.

Cooper looks down at Murph, who is standing at his elbow.

MURPH
What are you going to do with it?

COOPER
Reprogram it. Give it something socially responsible to do like drive a combine or a tractor.

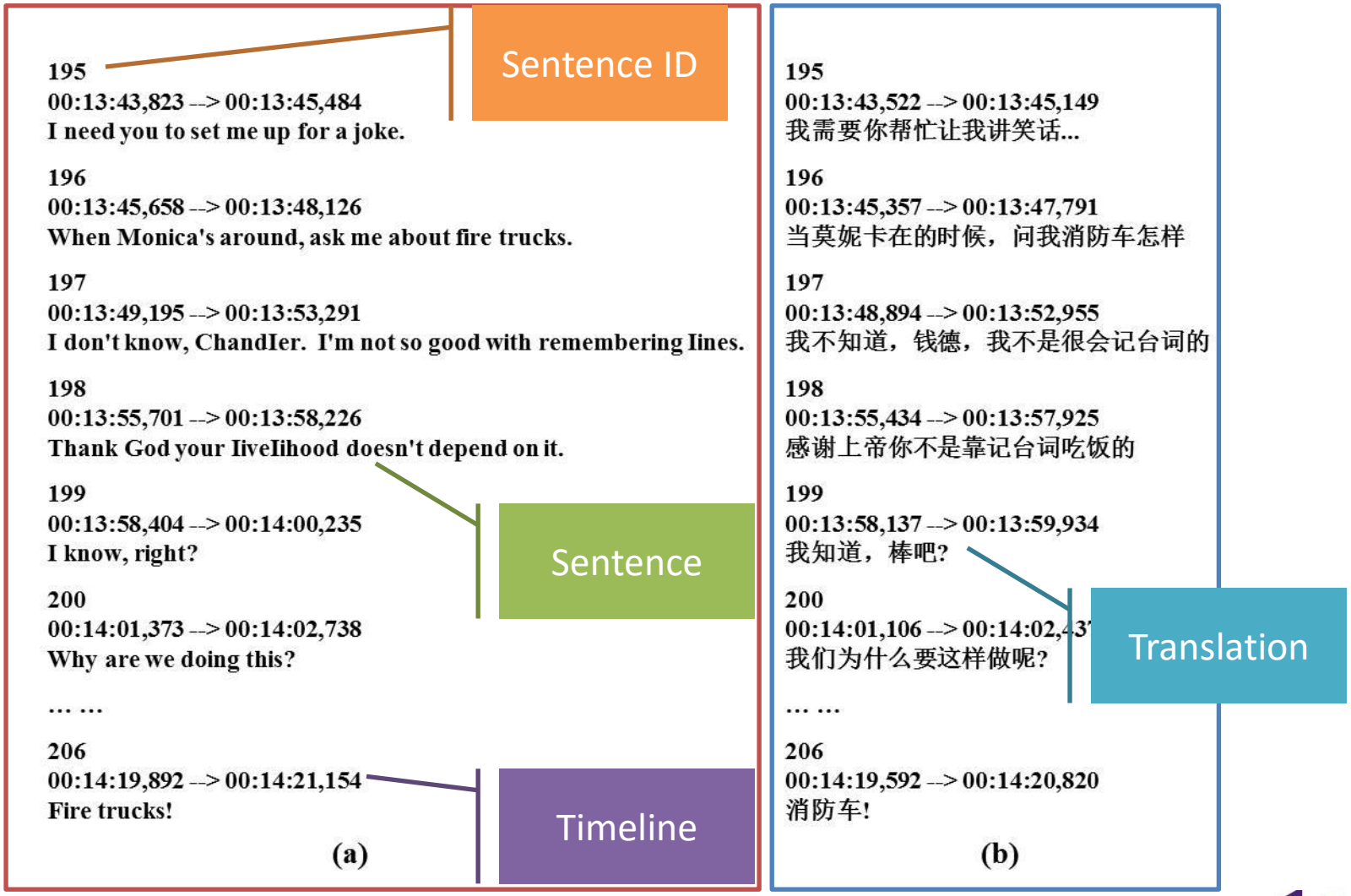
MURPH
(quiet)
Couldn't we just let it go? It's not hurting anyone.

Cooper looks down at his son. Good kid.

COOPER
We need all the help we can get,

Movie Scripts

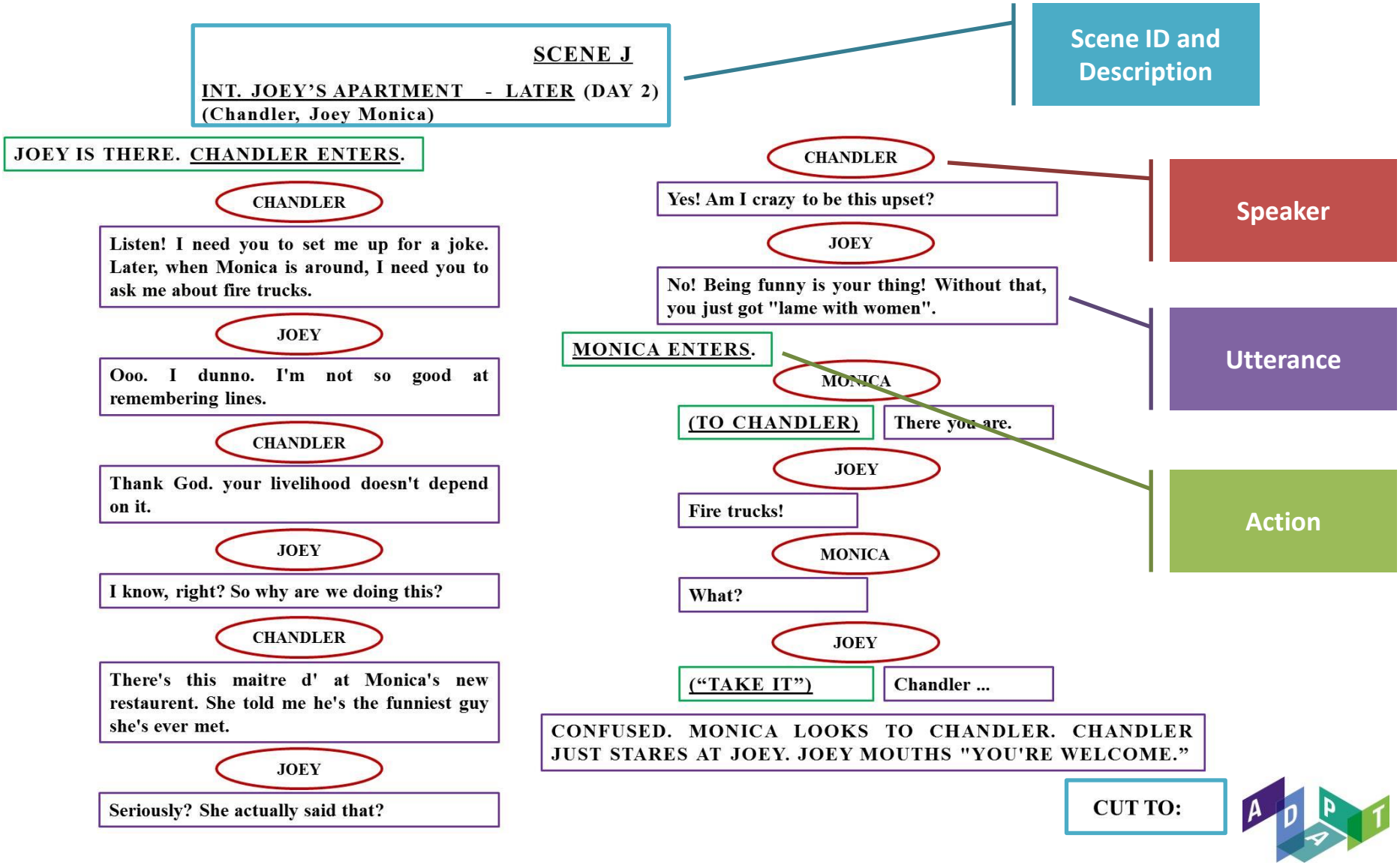
Sample of Movie Subtitles



English

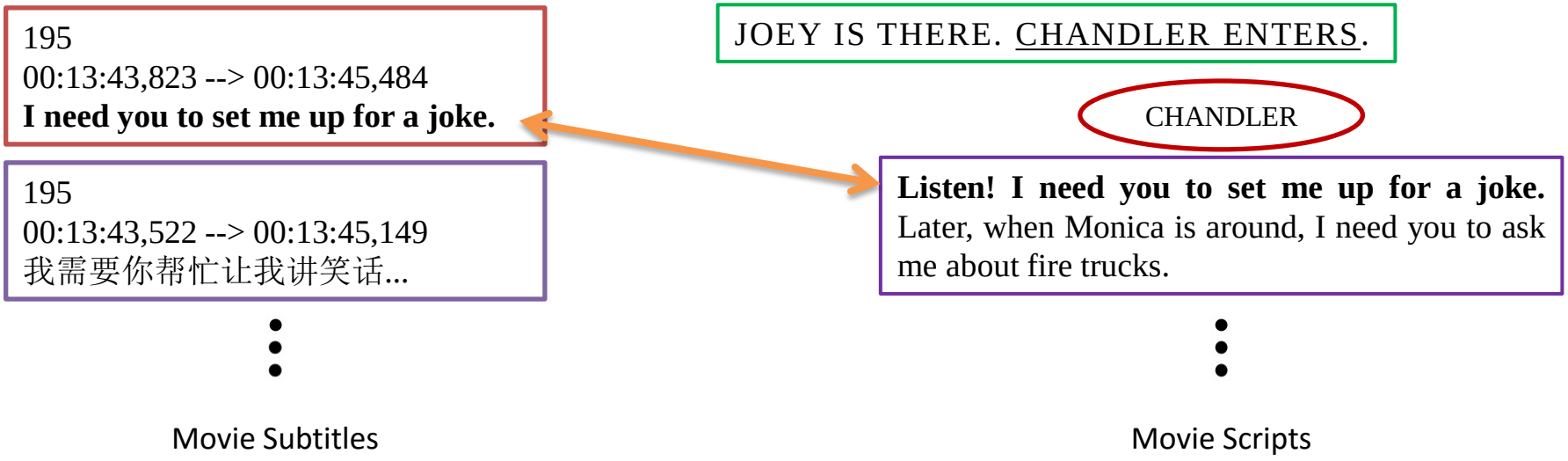
Chinese

Sample of Movie Scripts



Idea

- For the same movie, its subtitles and scripts always share the **same/similar** contents in the same language.



- This is a clue to align sentences between subtitles and scripts.
- Based on the alignment results, we can project the information from the script side to the subtitle side.
- **How about bridging these two kinds of resources?**

Proposed Approach

Automatic construction of dialogue corpus:

- Firstly, we extract parallel sentences from **bilingual subtitles**, and mine dialogue information from **monolingual movie scripts**.
- Secondly, we align sentences in between subtitles and scripts using **information retrieval (IR) approach**. We use each utterance in subtitle as a query to search the indexed script sentences.

$$sim(d_i, d_j) = \sum_{k=1}^{N} w_{i,k} \cdot w_{j,k} \sqrt{\sum_{k=1}^{N} w_{i,k}} \cdot \sqrt{\sum_{k=1}^{N} w_{j,k}} \quad (1)$$

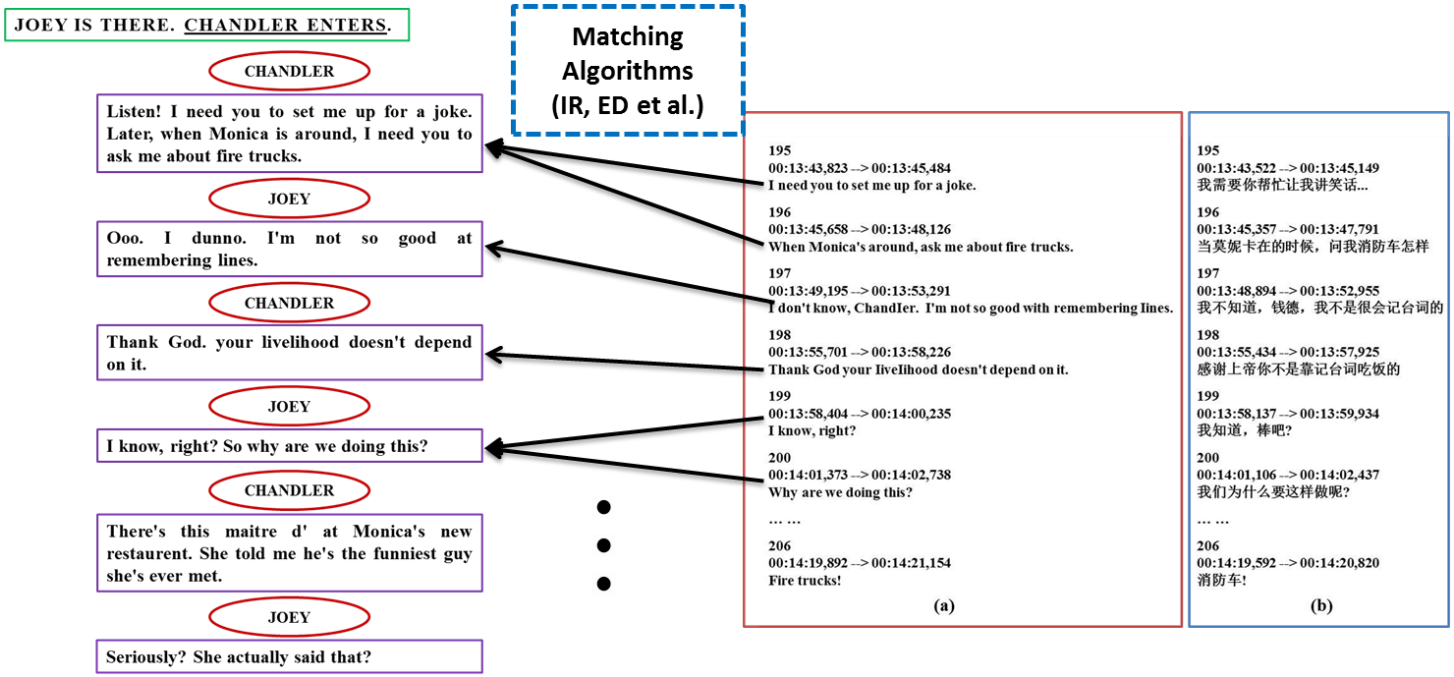
$$w_{t,d} = tf(t,d) \cdot idf(t,d,D) \quad (2)$$

- Thirdly, we **project** dialogue information (e.g. **speaker tag**, **scene boundary**, **action**) from the script side to the subtitle side.
- We can finally build parallel corpus with projected annotations.

Search and Projection

Inconsistency problems:

- many-to-many mapping (split into smallest units; combine and vote)
- variances in subtitles and scripts (stemmer, stop word and low case)
- short sentence and multiple occurrences (window)
- missing match (remove noise)



Projection Results

We conduct our experiments on the data extracted from the American TV play ***Friends***.

Applying the presented method, we obtain a Chinese–English dialogue corpus with projected information.

Item	Size
Total number of scripts processed	236
Total number of dialogues	5,428
Total number of speakers	42
Total number of utterances	109,268
Average amount of dialogues per script	23
Average amount of speakers per dialogue	3.5
Average amount of utterances per dialogue	20

Compared with gold standard reference (manually annotate), the agreements between automatic labels and manual labels is **81.79% on speaker** and **98.64% on dialogue boundary**, respectively.

Sample of Dialogue Corpus


```
<dialogue id="4884" n_utterances="12" scene= " JOEY'S APARTMENT - LATER (DAY 2) (Chandler, Joey, Monica) ">
  <context id="1" description= "JOEY IS THERE. CHANDLER ENTERS " >
    <utterance id="1" speaker="CHANDLER">
      <EN>I need you to set me up for a joke.</EN> <ZH>我需要你帮忙让我讲笑
    </utterance>
    <utterance id="2" speaker="CHANDLER">
      <EN>When Monica's around, ask me about fire trucks.</EN> <ZH>当莫妮卡在的时候，问我消防车怎样</ZH>
    </utterance>
    <utterance id="3" speaker="JOEY">
      <EN>I don't know, ChandIer. I'm not so good with remembering Iines.</EN> <ZH>我不知道，钱德，我不是很会记台词的</ZH>
    </utterance>
    <utterance id="4" speaker="CHANDLER">
      <EN>Thank God your IiveIihood doesn't depend on it.</EN> <ZH>感谢上帝你不是靠记台词吃饭的</ZH>
    </utterance>
    <utterance id="5" speaker="JOEY">
      <EN>I know, right?</EN> <ZH>我知道，棒吧?</ZH>
    </utterance>
    <utterance id="5" speaker=" JOEY ">
      <EN>W                                    > <ZH>我们为什么要这样做呢?</ZH>
    </utterance>
      … …
  </context>
  <context id="2" description = "MONICA ENTERS " >
    <utterance id="12" speaker= " MONICA" action= "TO CHANDLER" >
      <EN>Hi. There you are</EN> <ZH>嗨，你们都在</ZH>
    </utterance>
    … …
    <utterance id="12" speaker="JOEY">
      <EN>Fire trucks!</EN> <ZH>消防车!</ZH>
    </utterance>
  </context>
    … …
  <context id="3" description = "CONFUSED. MONICA LOOKS TO CHANDLER… …" >NULL</context>
</dialogue>
```

Sub-scene Description

Scene Description

Sentence & Translation

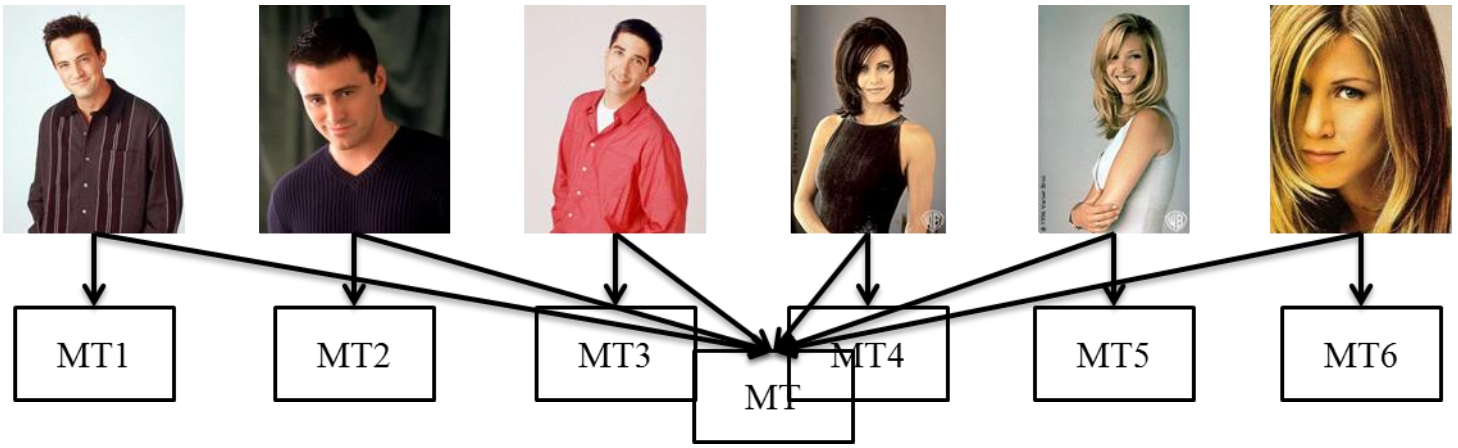
Scene Boundary

Speaker & Action

Machine Translation Experiment

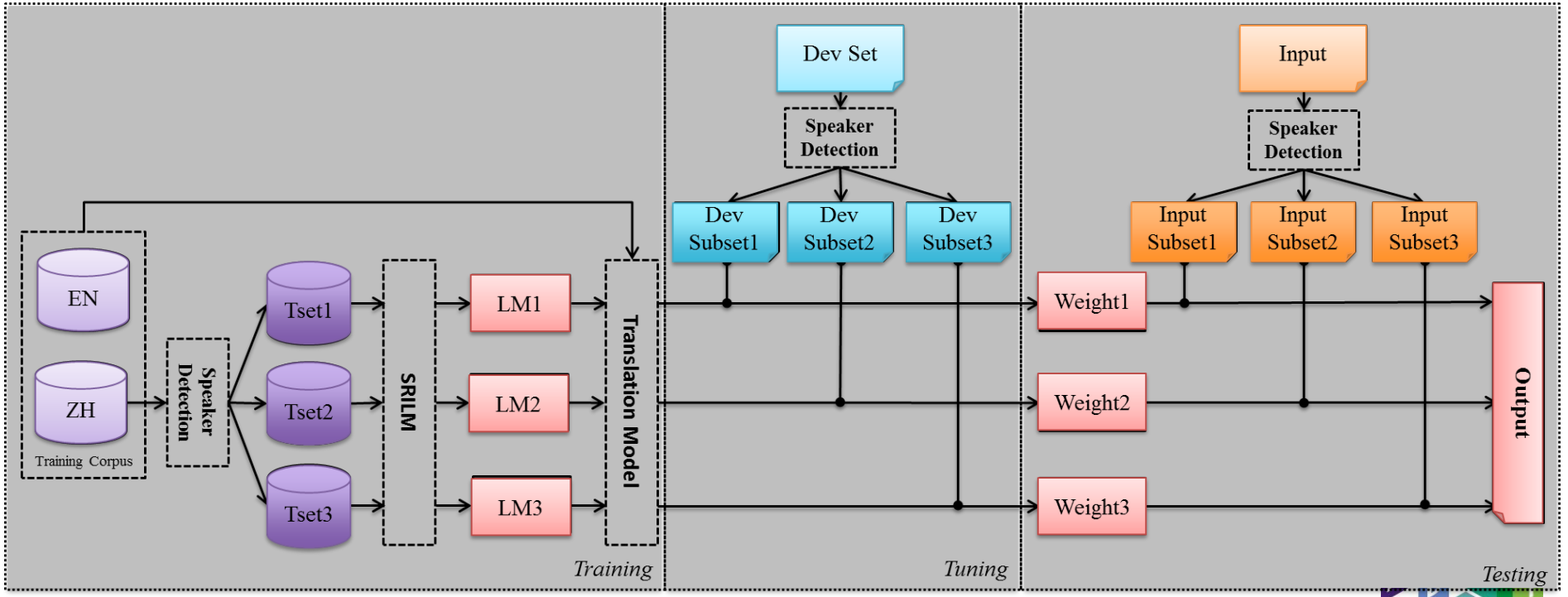
We preliminarily conduct an experiment to demonstrate how projected annotations (**speaker tags**) helps dialogue machine translation.

- persons in the movie have different roles, personal attributes (gender, age), backgrounds, characters etc.
- one person may have its specific language style, vocabulary, pet phrase etc.
- It is better to keep these hidden characteristics during translation.
- we build a personalized SMT system using the dialogue corpus.



Machine Translation Experiment

- **Language models** are trained on the target side of training corpus.
- Sentences in training, dev, test sets are split into ***N* subsets** according to the speaker tags (*N* = 7).
- **Tune** different parameter sets for each speaker-subset.
- **Decode** with parameter sets according to the speaker tags of inputs.



Machine Translation Results

The BLEU scores are low because only one reference and small-scale of training data.

For both directions, our method achieve **better results** than the baseline system.

- ZH-EN: it improves by **+0.87** BLEU score on test set
- EN-ZH: it improves by **+0.72** BLEU score on test set

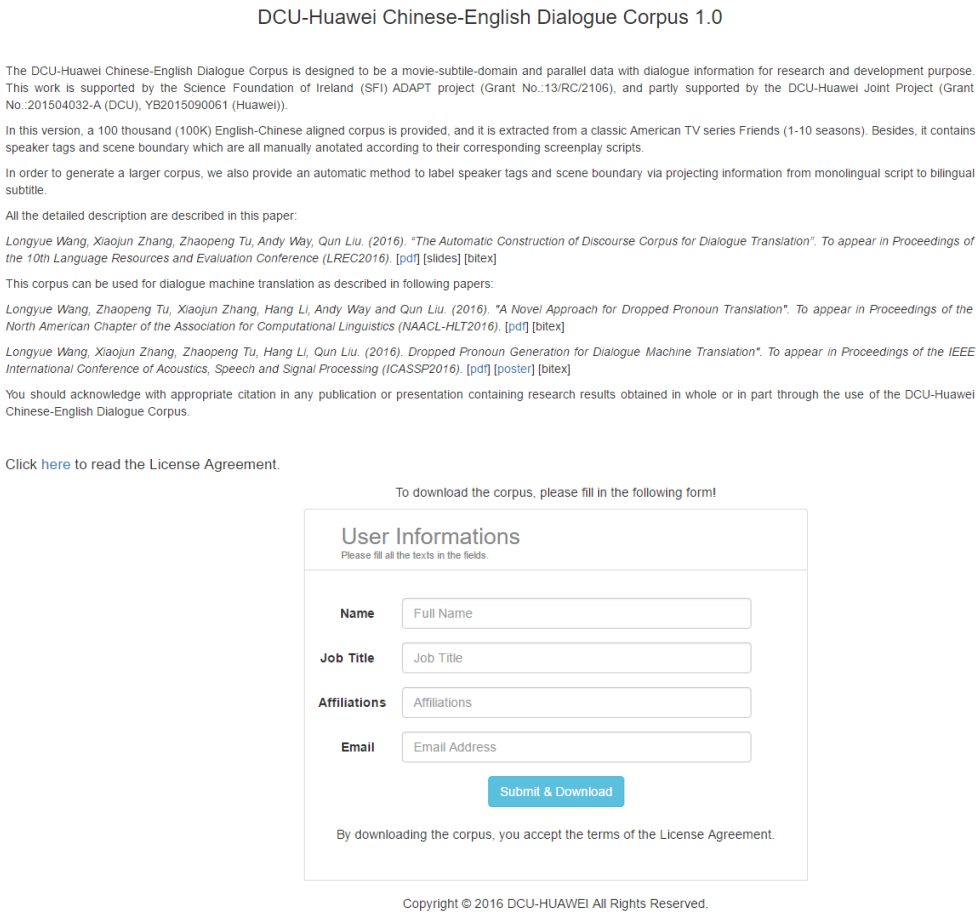
The results indicate that:

- the speaker tags can really help dialogue machine translation.
- our corpus construction method is relatively trustworthy.

System	Language Pair	Dev Set	Test Set
Baseline	ZH-EN	20.12	14.88
Personalized SMT	ZH-EN	22.01 (+1.89)	**15.75** (+0.87)
Baseline	EN-ZH	14.21	10.24
Personalized SMT	EN-ZH	16.05 (+1.84)	**10.96** (+0.72)

DCU-Huawei Chinese-English Dialogue Corpus 1.0

We also **manually** annotate the dialogue corpus based on **automatic results,** and release them in the website.



DCU-Huawei Chinese-English Dialogue Corpus 1.0

The DCU-Huawei Chinese-English Dialogue Corpus is designed to be a movie-subtitle-domain and parallel data with dialogue information for research and development purpose. This work is supported by the Science Foundation of Ireland (SFI) ADAPT project (Grant No.:13/RC/2106), and partly supported by the DCU-Huawei Joint Project (Grant No.:201504032-A (DCU), YB2015090061 (Huawei)).

In this version, a 100 thousand (100K) English-Chinese aligned corpus is provided, and it is extracted from a classic American TV series Friends (1-10 seasons). Besides, it contains speaker tags and scene boundary which are all manually anotated according to their corresponding screenplay scripts.

In order to generate a larger corpus, we also provide an automatic method to label speaker tags and scene boundary via projecting information from monolingual script to bilingual subtitle.

All the detailed description are described in this paper:

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Andy Way, Qun Liu. (2016). "The Automatic Construction of Discourse Corpus for Dialogue Translation". To appear in Proceedings of the 10th Language Resources and Evaluation Conference (LREC2016). [pdf] [slides] [bitex]

This corpus can be used for dialogue machine translation as described in following papers:

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way and Qun Liu. (2016). "A Novel Approach for Dropped Pronoun Translation". To appear in Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT2016). [pdf] [bitex]

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Hang Li, Qun Liu. (2016). Dropped Pronoun Generation for Dialogue Machine Translation". To appear in Proceedings of the IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP2016). [pdf] [poster] [bitex]

You should acknowledge with appropriate citation in any publication or presentation containing research results obtained in whole or in part through the use of the DCU-Huawei Chinese-English Dialogue Corpus.

Click here to read the License Agreement.

To download the corpus, please fill in the following form!

User Informations

Please fill all the texts in the fields.

Name: Full Name

Job Title: Job Title

Affiliations: Affiliations

Email: Email Address

Submit & Download

By downloading the corpus, you accept the terms of the License Agreement.

Copyright © 2016 DCU-HUAWEI All Rights Reserved.

Conclusion and Future Work

- We propose an approach to **build a parallel dialogue corpus** from monolingual scripts and their corresponding bilingual subtitles.
- We explore the effects of **speaker tags** on dialogue MT and it give positive results.
- Finally we release the DCU-Huawei English-Chinese Dialogue Corpus 1.0 at http://computing.dcu.ie/~lwang/corpora/resource.html.

In the future, we intend to:

- explore more information such as **scene boundary** in the dialogue corpus for translation tasks. **Longyue Wang**, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way and Qun Liu. 2016. "**A Novel Approach for Dropped Pronoun Translation**". in Proceedings of the NAACL-HLT2016 (long).
- build larger dialogue corpus using current resources such as OpenSubtitles2016 and IMSDb.

Thanks 謝謝

Longyue Wang 王龍躍
ADAPT Centre, Dublin City University
lwang@computing.dcu.ie

This work is supported by the Science Foundation of Ireland (SFI) ADAPT project (Grant No.:13/RC/2106), and partly supported by the DCU-Huawei Joint Project (Grant No.:201504032-A, YB2015090061).