Challenges of Multi-party Dialogue in Automatic Meeting Understanding

Raquel Fernández

Institute for Logic, Language & Computation
University of Amsterdam

www.illc.uva.nl/~raquel

UPC, June 2010
The CALO Meeting Assistant

**CALO: Cognitive Assistant that Learns and Organizes**

- Observes human-human meetings
  - Audio/video recording and speech/gesture recognition
  - Written/typed notes and paper/whiteboard sketches
- Produces a useful record of the interaction
The CALO Meeting Assistant

**CALO: Cognitive Assistant that Learns and Organizes**

- Observes human-human meetings
  - Audio/video recording and speech/gesture recognition
  - Writtenyped notes and paper/whiteboard sketches
- Produces a useful record of the interaction
- But understanding human-human spoken dialogue is hard
  - Informal, often ungrammatical language
  - Overlapping, fragmented speech
  - High speech recognition error rates (20-30% WER)
- Overhearing requires extra effort: can’t ask for clarification
- There is no point in trying to understand everything: instead, target some conversational events that we can understand
Overview

• Detecting Decisions
  – Exploiting interaction structure is crucial

• Resolving personal pronouns (‘you’)
  – Exploiting contextual and visual information is crucial

○ All joint work with:
  – Matt Purver at Quen Mary University of London
  – Matthew Frampton, Patrick Ehlen, Stanley Peters at CSLI, Stanford
Detecting Decisions

First attempt to detect decision discussions automatically is due to Hsueh & Moore (2007). Using the AMI Meeting Corpus:

- Mark decision-related utterances, based on whether they belong to an extractive summary
- Binary classification (*decision dialogue act*: yes/no) using MaxEnt classifier
- Poor performance: ca. 35% F-score

What seems to be the problem?
- Decision-related utterances for a very heterogeneous set
  - they perform different dialogue functions,
  - and are related to different components of the decision
- Discussions tend to be split across utterances and participants
Decision Discussions

• Our approach, following Purver’s et al. work on action items, is to see decision discussion as a type of social action
  – with internal structure: multiple participants making different types of contributions

• What are the main components of decision discussions?
  – Decision-making dialogue can be complex, with many relevant components: alternative proposals, arguments for/against them, agreements and disagreements, ...

• Three main components that often unfold in sequence
  – a topic or issue is initially raised
  – one or more proposals are considered
  – once agreement is reached, a decision is adopted
Exploiting Discourse Structure

- Different decision dialogue acts play different roles

1a: Are we going to have a backup? Or we do just–
2b: But would a backup really be necessary?

RP 3a: I think maybe we could just go for the kinetic energy and be bold and innovative.

A 4c: Yeah.
A 5b: I think– yeah.
6a: It could even be one of our selling points.
7c: Yeah <laugh>
8d: Environmentally conscious or something.
9a: Yeah.

A RR 10b: Okay, fully kinetic energy.
A 11d: Good.

- These sub-classes may be more homogeneous than looking for just decision-related dialogue acts
- A more fine-grained distinction at the detection stage can help us extract summaries of decision discussions
Corpus & Annotation

- Annotation scheme for decision dialogue acts (DDAs):

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>issue</td>
</tr>
<tr>
<td>R</td>
<td>resolution</td>
</tr>
<tr>
<td>RP</td>
<td>proposal</td>
</tr>
<tr>
<td>RR</td>
<td>restatement</td>
</tr>
<tr>
<td>A</td>
<td>agreement</td>
</tr>
</tbody>
</table>

- 17 scenario-based dialogues from the AMI Meeting Corpus
- ~ 30 min/dialogue, 15,680 utts
- 671 annotated DDAs (4.3% of utts)
- 2 annotators, $\kappa = .63 - .73$
- Very small overlap with AMI annotations: 12% of our DDAs
Hierarchical Classification

Hierarchical approach involving two steps:

- **DDA identification**: four independent *sub-classifiers*
  - Features: words, utterance length/duration, pitch & intensity, AMI DAs, speaker role, context

- **Decision Discussion detection**: *super-classifier*
  - Features: output of sub-classifiers over a window of N utterances
  - The super-classifier is able to “correct” the hypothesized DDA

- Support Vector Machines, 17-fold cross validation
Decision Detection Results

- Results for individual DDA are rather low: 15-40% F-score
- For Decision Discussion detection, the super-classifier improves over “flat” approach
  - Flat baseline: 45% F-score
  - Hierarchical: recall 88%, precision 43%, F-score 63%
- Taking into account Agreement helps significantly: 52% → 63%
- Robustness to ASR output is reasonable: sub-classifier performance degrades, but super-classifier seems rather robust (7% F-score drop)
  - New approach by Bui & Peters (to appear): Hierarchical directed graphical model, F-scores 70-80%

Personal References in Decisions

- Some decisions like *action items* involve assigning actions to someone (the *owner* of that action item).
- Sometimes people use names, but only < 5% of cases. It’s much more common to volunteer yourself (‘*I’ll do X*’) or suggest someone else (‘*Maybe you could*’)
  - Heavy use of personal pronouns
  - ‘*I*’ → speaker identification problem
  - ‘*you*’ → addressee identification problem

- Addressee identification is hard, but approachable: Katzenmaier et al., 2004; Jovanovic et al., 2006: 70-80% accuracy.
- However, resolving ‘*you*’ is not always a matter of identifying the addressee....

⇒ ‘*You*’ resolution
Overview

- **Detecting Decisions**
  - Exploiting interaction structure is crucial

- **Resolving personal pronouns** (*you*)
  - Exploiting contextual and visual information is crucial

  - All joint work with:
    - Matt Purver at Quen Mary University of London
    - Matthew Frampton, Patrick Ehlen, Stanley Peters at the Computational Semantics Lab, Stanford.
Second-person English Pronouns: ‘You’

- ‘You’ as deictic – it refers to the addressee.
- This is trivial in two-party dialogue, but not so in multi-party...
- In fact, it’s not so trivial in two-party dialogue either since ‘you’ is often not a referring pronoun at all

A: Well you know I think we need to get the application in
B: Uh-huh
A: As soon as possible really
B: Yeah if you get it in by November you get a discount
A: Yeah so do you think you could do that?
C: OK sure
A: Great. Do you guys have any further questions?
Second-person English Pronouns: ‘You’

- Generic pronouns refer to no-one in particular
- Discourse markers are not referential

(1) Often **you** need to know specific button sequences to get certain functionalities done.
(2) It’s not just, **you** know, noises like something hitting.

- Deictic uses refer to the addressee – but how many? only one? more than one? everyone in the audience or just a subset?

(3) John, I think **you** should take care of the reservations
(4) John and Pat, could **you** do your presentations tomorrow?
(5) Do **you** guys have any further questions?
(6) Are **you** all ready?
Data and Categories

- We use the AMI Meeting Corpus (McCowan et al, 2005)
- 948 you utterances from 10 different meetings, annotated as either Discourse marker, Generic, or Deictic
- We use the AMI addressee annotations for deictic uses, and collapse 2- and 3-participant addressees since they make up >2%

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discourse marker</td>
<td>72</td>
<td>8%</td>
</tr>
<tr>
<td>Generic</td>
<td>431</td>
<td>45%</td>
</tr>
<tr>
<td>Deictic</td>
<td>445</td>
<td>47%</td>
</tr>
<tr>
<td><strong>singular</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>plural</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>pl.all</strong></td>
<td>290</td>
<td>68%</td>
</tr>
<tr>
<td><strong>pl.subgroup</strong></td>
<td>137</td>
<td>32%</td>
</tr>
</tbody>
</table>

- Discourse markers are easy to identify, so we ignore them 
  ¬(do | as) you know ¬(how | that) → 99% accuracy
To classify the referent of singular deictic instances, we use two labelling schemes:

- **Position**: identifies the potential addressees by their seating position relative to the current speaker: opposite, diagonal, next to
- **Roles**: identifies participants according to their meeting role: industrial designer (ID), marketing expert (ME), project manager (PM), and user-interface designer (UI).

<table>
<thead>
<tr>
<th>Total</th>
<th>Position rel. to speaker</th>
<th>Individual participant roles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1</td>
<td>L2</td>
</tr>
<tr>
<td>100%</td>
<td>35%</td>
<td>30%</td>
</tr>
<tr>
<td>290</td>
<td>102</td>
<td>88</td>
</tr>
</tbody>
</table>

- ‘You’-resolution as a 5-way or 6-way problem:
  - Generic, L1, L2, L3, Plural
  - Generic, ID, ME, PM, UI, Plural
Sources of Information

Multiple sources of information available:

• Transcripts (manual and ASR)
• Speech / prosody (close-talking mics)
• Dialogue act annotations (manual)
• Video (manual annotation and automatic extraction)

Different types of features:

• Intra-utterance
• Contextual: across utterances
• Visual: extra-linguistic
Intra-utterance Features

- **Sentential** features encoding lexical and shallow syntactic patterns:
  - ‘if you’, ‘often’, ... ⇝ generic?
  - ‘do you’, ‘you said’, ... ⇝ deictic?
  - ‘you guys’... ⇝ plural?
  - participant Named Entity tag ⇝ singular addressee?

- **Dialogue act** features from the AMI DA annotations
  - questions, commands ⇝ deictic?

- **Prosodic** features such as pitch and intensity, extracted with Praat
  - ‘you’ more stressed ⇝ deictic?

- **Lexical** features
  - words/n-grams seen more than once in the corpus
Contextual Features

Features extracted from pairs of utterances, forward- and backward-looking:

- **Speaker activity**
  - who speaks next and spoke last?
  - how many utterances in between? how many different speakers?

- **Utterance context**
  - overlaps, long pauses between utterances $\leadsto$ not addressee?
  - ratio of common words $\leadsto$ addressee?
  - Dialogue act combinations $\leadsto$ addressee?
Visual Features

- **Who’s the speaker looking at?**
  - when ‘you’ is uttered
  - at the beginning / middle / end of the utterance

- **Who are the listeners looking at?**
  - when ‘you’ is uttered
  - at the beginning / middle / end of the utterance

- **Is there mutual gaze between the speaker and a listener?**
  - when ‘you’ is uttered
  - at the beginning / middle / end of the utterance
Visual Features

Visual features also extracted automatically with a head-tracker over the video sequence (by M. Christoudias & T. Darrell at ICSI, UC Berkeley)
Learning to Resolve ‘You’

• Technicalities:
  – We use a Bayesian Network classifier for all experiments
  – All results after 10-fold cross-validation

• Approached as a 5/6-way problem, performance is fairly poor:
  – Majority class baseline (always generic): 51% accuracy
  – Best classifier after feature selection: 62% acc. (56% with ASR and automatic features)
  – F-scores: 75% generic, 41% plural, 38-60% individual addressees

• Different kinds of problems?
  – Some aspects seem intra-sentential: generic / deictic distinction
  – Some aspects seem interactional: individual addressee reference
  – Some aspects might be in between: singular / plural distinction
### Generic vs. Deictic you

<table>
<thead>
<tr>
<th></th>
<th>Acc.</th>
<th>F1-Gen</th>
<th>F1-Deic</th>
<th>Acc.ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC Baseline</td>
<td>50.80</td>
<td>0</td>
<td>.67</td>
<td>53.51</td>
</tr>
<tr>
<td>Best classifier</td>
<td>86.99</td>
<td>.87</td>
<td>.87</td>
<td>84.58</td>
</tr>
</tbody>
</table>

- Intra-utterance features are the best predictors.
  - Generic words ‘always’, etc.; multiple ‘you’ s – generic names, 1st person pronouns – deictic
  - DAs help: questions are more likely to contain deictic ‘you’.
  - Lexical features are very useful (79% → 87%):
    - meeting main topic (*button, channel, volume*) – generic uses
    - meeting management (*presentation, email, project, meeting*) – deictic uses

- Contextual and visual features do not help.
Singular vs. Plural Deictic you

<table>
<thead>
<tr>
<th></th>
<th>Acc.</th>
<th>F1-Sing</th>
<th>F1-Plural</th>
<th>Acc.ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC Baseline</td>
<td>67.92</td>
<td>.81</td>
<td>0</td>
<td>65.38</td>
</tr>
<tr>
<td>Linguistic features</td>
<td>83.37</td>
<td>.88</td>
<td>.72</td>
<td>77.35</td>
</tr>
</tbody>
</table>

- **Intra-utterance features are the most useful ones**
  - Questions tend to be addressed to singular participants.
  - Plural 1st person pronouns (‘we’) - plural ‘you’.

- **Contextual features are also good predictors**
  - Utterance similarity - singular addressee
  - Speaker activity: one speaker - plural / more interaction - singular

- **Visual features don’t help at all (66% accuracy)**
  - although gaze at whiteboard correlates with plural
Singular Addressee Identification

<table>
<thead>
<tr>
<th></th>
<th>Acc.</th>
<th>F1-$L_1$</th>
<th>F1-$L_2$</th>
<th>F1-$L_3$</th>
<th>Acc.ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC Baseline</td>
<td>35.17</td>
<td>.52</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>NS Baseline</td>
<td>71.03</td>
<td>.68</td>
<td>.74</td>
<td>.71</td>
<td></td>
</tr>
<tr>
<td>Linguistic features</td>
<td>72.07</td>
<td>.72</td>
<td>.73</td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>Visual</td>
<td>74.14</td>
<td>.81</td>
<td>.69</td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>Multimodal</td>
<td><strong>84.14</strong></td>
<td>.86</td>
<td>.84</td>
<td>.82</td>
<td><strong>74.14</strong></td>
</tr>
</tbody>
</table>

- Intra-sentential features alone yield poor results (49% acc.)
  - Dialogue acts are not predictive in this case
- Contextual features are very good predictors
  - Next speaker “baseline”: 71%
  - Lexical similarity, time separation, distance in utterances
  - Backward-looking only (online): 59%
- Visual features are excellent predictors (74%)
  - The two labelling schemes differ:
    - Roles: captures lexical information related to participant roles (*sales, make money* → marketing expert; *speech rec, technical* → industrial designer)
    - Positions: better suited to encode visual information.
In Progress / Future Work

- Cascaded classifier:
  generic vs. deictic → plural vs. singular → individual addressees
  - Use optimal features (utterance, context, visual) for each problem
  - 78% accuracy (72% with ASR and automatic features)

- Future work: Distributional approach for generic vs. deictic, inspired by Bergsma et al. (2008)’s work on identifying non-referential “it”

<table>
<thead>
<tr>
<th>Original</th>
<th>Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>You can make it in advance</td>
<td>make * in advance</td>
</tr>
<tr>
<td>You can make it in Hollywood</td>
<td>make * in Hollywood</td>
</tr>
</tbody>
</table>

Several NP fillers ⇒ referential / “it” filler ⇒ non-referential