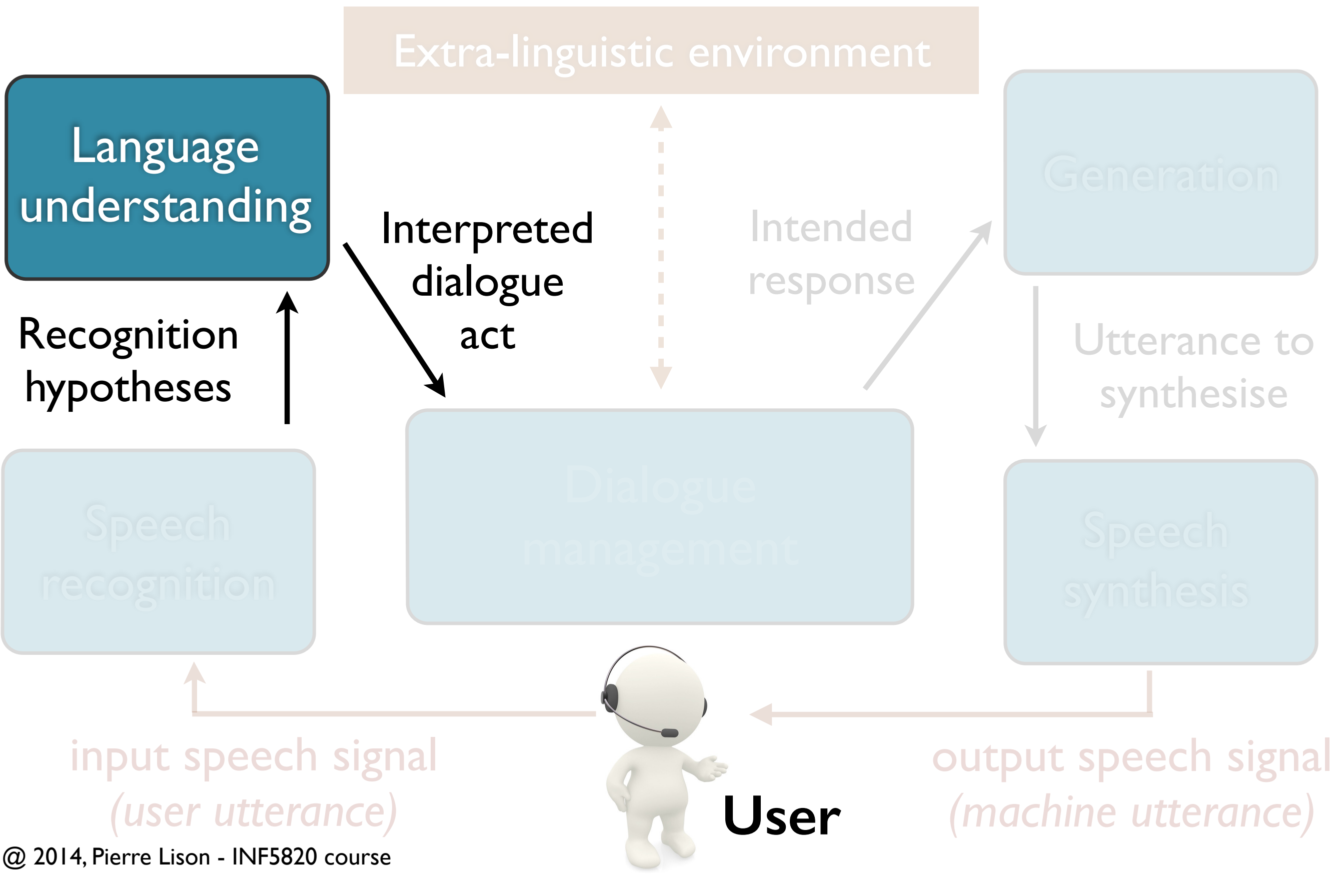
UiO : University of Oslo

INF5820: Language Understanding

Pierre Lison,
Language Technology Group (LTG)
Department of Informatics

Fall 2014

Language understanding



Outline

- Parsing spoken language
- Three challenges
- Reference resolution
- Dialogue act recognition
- Summary

UiO : University of Oslo

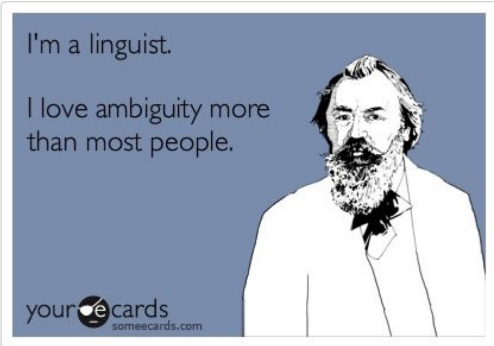
Outline

- **Parsing spoken language**
- Three challenges
- Reference resolution
- Dialogue act recognition
- Summary

Parsing

- The goal of parsing is to build a representation of the *meaning*(s) expressed by the *form* of a given utterance on the basis of its *grammatical structure*

- Major challenges:
  - Coverage
  - Robustness
  - Ambiguity resolution



Parsing

- Common approaches:
  - **Shallow parsing** (e.g. concept spotting): small task-specific patterns used to extract specific constituents and turn these into basic semantic concepts
  - **Grammar-based parsing**: generic grammars (possibly adapted to spoken dialogue) used to extract possible syntactic relations
  - **Statistical parsing**: probabilistic models of syntactic structure trained on spoken data

Parsing

	Pros	Cons
Shallow parsing	• Efficient • Easy to understand and develop • Direct mapping to domain-specific semantics	• Domain-specific • Manual development effort • Limited coverage
Grammar-based parsing	• Reusable grammar • Yields more fine-grained structures than shallow parsing	• Grammar rules must be adapted/relaxed for spoken dialogue • Limited coverage & robustness • Parse selection problem • Efficiency concerns
Statistical parsing	• Increased robustness • Learning algorithm is reusable	• Requires training data! • Difficult to model sophisticated linguistic phenomena

Shallow parsing

- Most popular approach in current spoken dialogue systems
- Concentrate on specific *information-bearing phrases*, ignoring the rest
  - Example: locative phrases and temporal expressions for a flight-booking system

```
<top> =  leaving for <city>
         fly to <city> (<time>)
         fly from <city>
         departing from <city>
         arrive before <time>
         ...
```

```
<time> = on <month> <date>
         at <hour>
         tomorrow
         ...
<city> = Los Angeles
         Oslo
         Madrid
         ...
```

Shallow parsing

- Advantages of shallow parsing:
  - Direct mapping to the set of semantic concepts that are relevant for the application at hand
  - If the ASR language model is anyway constrained by a grammar, shallow parsing is the best option
- But can lead to robustness & coverage problems for more complex language models

[Dowding et al. (1994). «Interleaving syntax and semantics in an efficient bottom-up parser.» In *ACL-94*]

[J. Allen et al (1996). «A robust system for natural spoken dialogue». In *ACL'96*]

Grammar-based parsing

- Alternative: perform a real *grammatical analysis*
  - Outputs the set of possible analyses for the utterance
  - Domain-independent grammar
- Challenges:
  - *Coverage and robustness* against disfluencies, non-sentential utterances and ASR errors (need to *relax* rules)
  - Must be followed by a parse selection step (disambiguation)

[G. van Noord (1999). «Robust grammatical analysis for spoken dialogue systems». *Journal of Natural Language Engineering]*

Statistical parsing

- Third approach: *train* a parser directly from data
  - Flat models (HMM tagging of semantic concepts)
  - Structured models (PCFGs, transition-based parsing, etc.)
- Advantages:
  - Improved coverage & robustness
  - Direct selection of most likely parse(s)
- Major concern: for most applications, data is *scarce*, expensive to acquire, and highly domain-specific

[He, Y. and Young, S. (2005). «Semantic processing using the Hidden Vector State Model», in *Computer Speech and Language*]

Outline

- Parsing spoken language
- **Three challenges:**
  - **Speech recognition errors**
  - **Disfluencies**
  - **Non-sentential utterances**
- Reference resolution
- Dialogue act recognition
- Summary

Speech recognition errors

- Speech recognition errors are pervasive
  - often between 15-25 % in a normal dialogue domain
- It has been shown that post-processing the ASR output can improve accuracy
  - Can be trained given annotated data (speech recognition output associated with gold-standard transcription)
  - Noisy-channel model to represent the (probabilistic) relation between the actual and intended output

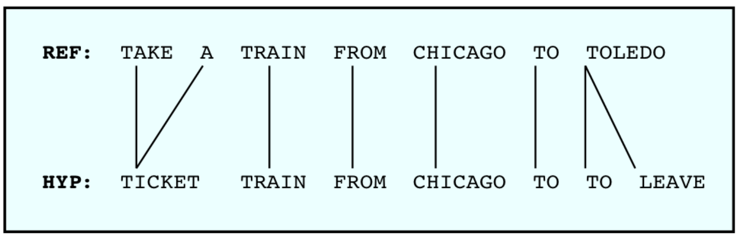
[E. Ringger & J. Allen (1996), «Error correction via a post-processor for continuous speech recognition», in *ICASSP'96*]

Speech recognition errors

- Noisy-channel model: given an ASR hypothesis Y, find the most likely original sentence X:

$$\hat{X} = \underset{X}{\operatorname{argmax}} \Pr(Y|X) \Pr(X)$$

  - *Language model* P(X): describes the prior probability of utterance X
  - *Channel model* P(Y|X): describes the most likely confusions X → Y realised by the speech recogniser («fertility model» allowing n-to-m mappings)



Disfluencies

- Speakers construct their utterances «as they go», incrementally
  - Production leaves a *trace* in the speech stream
- Presence of multiple disfluencies
  - Pauses, fillers («øh», «um», «liksom»)
  - Fragments
  - repetitions («the the ball»), corrections («the ball err mug»), repairs («the bu/ ball»)

Disfluencies

- Internal structure of a disfluency:

$$\text{Book a ticket } \underbrace{\text{to Boston}}_{\text{reparandum}} \ \underbrace{\text{uh I mean}}_{\text{interregnum}} \ \underbrace{\text{to Denver}}_{\text{repair}}$$

- reparandum: part of the utterance which is edited out
- interregnum: (optional) filler
- repair: part meant to replace the reparandum

[Shriberg (1994), «Preliminaries to a Theory of Speech Disfluencies», Ph.D thesis]

Basic examples of disfluencies

- Repetitions

$$\text{robot now go to } \underbrace{\text{the hallway}}_{\text{reparandum}} \underbrace{\text{the hallway}}_{\text{repair}}$$

- Corrections:

$$\text{ok and then turn } \underbrace{\text{right}}_{\text{reparandum}} \underbrace{\text{no sorry I mean}}_{\text{interregnum}} \underbrace{\text{left}}_{\text{repair}}$$

- Rephrasing/completion:

$$\text{robot please give me } \underbrace{\text{the ball}}_{\text{reparandum}} \underbrace{\text{yes}}_{\text{interregnum}} \underbrace{\text{the red one on your left}}_{\text{repair}} \text{ exactly}$$

Remarks on disfluencies

- All parts of a disfluency may carry *meaning* relevant for interpretation
  - Even filled pauses such as «uh» and «um»
- The syntactic types of the reparandum and repair need not be identical (ex: "turn to the left err no forward")
  - Levelt: reparandum and repair are of syntactic types that *could* be joined by a conjunction
- Pervasive phenomena: about 6% of the words in spontaneous speech are «edited»

[Levelt W. (1983), « Monitoring and self-repair in speech», in *Cognitive Science*.]

More complex disfluencies

så <u>gikk jeg</u> e <u>flytta vi</u> til Nesøya da begynte jeg på barneskolen der

og så har jeg gått på Landøya ungdomsskole # som ligger ## <u>rett over broa nesten</u> # <u>rett med Holmen</u>

jeg gikk på Bryn e skole som lå rett ved der vi bodde den gangen e <u>barneskole</u>

videre på Hauger ungdomsskole

da <u>hadde alle hele på skolen skulle</u> liksom # spise julegrøt og <u>det va- det var</u> bare en mandel

og da var jeg som fikk den da ble skikkelig sånn " wow # jeg har fått den " ble så glad

[*«Norske talespråkskorpus - Oslo delen»* (NoTa), collected and annotated by the Tekstlaboratoriet]

Treatment of disfluencies

- Motivation: words in reparandum usually closely related to those in the repair
- Given observed sentence Y, search for:

$$\hat{X} = \underset{X}{\operatorname{argmax}} \Pr(Y|X) \Pr(X)$$

  - Language model $\Pr(X)$: bigram, trigram, syntax-based
  - Channel model $\Pr(Y|X)$: TAG matching reparandum to repair using deletion, insertion, substitution.

[Johnson, M. & Charniak, E. «A TAG-based noisy channel model of speech repairs», Proceedings of ACL 2004]

Treatment of disfluencies

- Previous research mostly targeted on disfluency detection in *human-human* dialogues (e.g. Switchboard)
- Less work on the treatment of disfluencies in *human-machine* dialogues
  - **Easier**: less disfluencies in human-machine dialogues (human users adapt to the machine), and some pre-filtering is already made by the speech recogniser itself
  - **More difficult**: need to work on real ASR outputs instead of gold-standard transcripts

Treatment of disfluencies

- Most papers on disfluencies assume that these can simply be *removed* from the input
- But disfluencies can contain important semantic information!
  - Example: «take the red ball uh yes the one to your left»
- Open research question: can we integrate disfluencies *as part of the grammatical analysis*, instead of simply filtering them out?

Paradigmatic piles

- Concept of "paradigmatic piles" in linguistics:
  - Paradigmatic pile = position in a utterance where the same syntactic position is occupied by several entities
  - Non-functional relations between phrases
  - Piles viewed as a *complement* to dependency relations (syntax expressed as a two-dimensional structure)
  - *Descriptive account* of phenomena such as disfluencies, reformulation, appositions, coordinations, etc.
  - Represented in a grid

[Benveniste, C.-B. (1998), «*Le francais parlé: études grammaticales*», Éd. du CNRS]

Disfluency and coordination

(a) Felix is a linguist, maybe a computer scientist [Disfl]
(b) Felix is a linguist uh maybe a computer scientist [Disfl]
(c) Felix is a linguist or maybe a computer scientist [Coord]
(d) Felix is a linguist and maybe a computer scientist. [Coord]

- (c) has the same interpretation as (b)
- (a) can either be interpreted «disjunctively» as in (b),(c), or «additively» as in (d)
- The syntactic types accepted in disfluencies and in coordination are similar (cf. Levelt's rule)

[Gerdes K., Kahane S. (2009), «Speaking in piles: Paradigmatic annotation of French spoken corpus», Processing of the 5th Corpus Linguistics Conference]

Disfluency and coordination (2)

(a)	Felix is		a linguist
		maybe	a computer scientist
(b)	Felix is		a linguist
		uh maybe	a computer scientist
(c)	Felix is		a linguist
		or maybe	a computer scientist
(d)	Felix is		a linguist
		and maybe	a computer scientist.

- Paradigmatic piles provide an unified treatment of (a)-(d)
- «maybe», «and» etc. are are *pile markers*
  - Pile structure similar for the 4 examples, but the final interpretation slightly different due to the distinct markers

Example of grid analysis

vokst opp i et stort				
	stort hus med		tre etasjer	
		og	mange rom i hver etasje	
		og	store rom	
			god plass	
			lun e	
			lun e	
			sånn gårdsstemning	i hvert rom ja
og				
ja				
nå bor jeg jo i en mer	urban			
	minimalistisk			
	moderne		leilighet	

NB: elegant, but purely descriptive account (no formal, computational treatment)

[«*Norske talespråkskorpus - Oslo delen*» (NoTa), collected and annotated by the Tekstlaboratoriet]

Non-sentential utterances

- **Non-sentential utterances** are utterances that lack an overt predicate
  - Pervasive: ± 30% of utterances, depending on the corpus
- Examples:
  - «Should I take the ball?» → *«yes indeed»*
  - «Please go the kitchen» → *«go where?»*
  - «Task completed» → *«brilliant!»*
  - «First take left after the corner» → *«and afterwards?»*

[J. Ginzburg (2012), «*The Interactive Stance: Meaning for Conversation*», OUP]

UiO : University of Oslo

Non-sentential utterances

- The meaning of non-sentential utterances is (practically always) *context-dependent*
  - Their meaning arises through the interaction itself
- This can lead to ambiguities in the resolution:

1. A: When are they going to open the new main station?
   B: Tomorrow (short answer)
2. A: They are going to open the station today.
   B: Tomorrow (correction)
3. A: They are going to open the station tomorrow.
   B: Tomorrow (acknowledgement)

[R. Fernandez (2006), «*Non sentential utterances in dialogue: classification, resolution and use*», PhD thesis]

Non-sentential utterances

- Classifying NSUs can be done with classical machine learning techniques
- About 20 classes can be reliably annotated
- Standard morpho-syntactic features: part-of-speech tags, presence of certain words, etc.
- Classification accuracy around 81%

[R. Fernández, J. Ginzburg, S. Lappin (2007), «Classifying Non-Sentential Utterances in Dialogue: A Machine Learning Approach», *Computational Linguistics*]

Non-sentential utterances

NSU class	Example	Total
Plain Acknowledgement	*A: ... B: mmh*	599
Short Answer	*A: Who left? B: Bo*	188
Affirmative Answer	*A: Did Bo leave? B: Yes*	105
Repeated Ack.	*A: Did Bo leave? B: Bo, hmm.*	86
C(larification) E(llipsis)	*A: Did Bo leave? B: Bo?*	79
Rejection	*A: Did Bo leave? B: No.*	49
Factive Modifier	*A: Bo left. B: Great!*	27
Repeated Aff. Ans.	*A: Did Bo leave? B: Bo, yes.*	26
Helpful Rejection	*A: Did Bo leave? B: No, Max.*	24
Sluice	*A: Someone left. B: Who?*	24
Check Question	*A: Bo isn't here. Okay?*	22
Filler	*A: Did Bo ... B: leave?*	18
Bare Mod. Phrase	*A: Max left. B: Yesterday.*	15
Propositional Modifier	*A: Did Bo leave? B: Maybe.*	11
Conjunction + frag	*A: Bo left. B: And Max.*	10
Total dataset		**1283**

[J. Ginzburg (2012), «*The Interactive Stance: Meaning for Conversation*», OUP]

Non-sentential utterances

- *Interpreting* NSUS is much trickier
  - **Sententialist approach**: view NSUs as the reduced form of an original, well-formed sentence
  - **Constructionalist approach**: NSUs are incorporated in the grammar as distinct constructions which specify a.o. the contextual characteristics which govern their use
- Some work done in the area of formal semantics, but lack of practical, real-scale implementations

[D. Schlangen and A. Lascarides, «The interpretation of non-sentential utterances in dialogue», in *SIGDIAL* 2003]
[J. Ginzburg (2012), «*The Interactive Stance: Meaning for Conversation*», OUP]

Outline

- Parsing spoken language
- Three challenges
- **Reference resolution**
- Dialogue act recognition
- Summary

Reference resolution

- **Reference resolution** is the process of finding which *entities* are *referred to* by specific linguistic expressions
  - Related to the notion of *deictics* (lecture 2)
  - The entity can be an object, a person, an event, a concept, etc.
- Complex problem in discourse and dialogue processing (we'll only scratch the surface here)

«*This presentation* was written yesterday»

«There's *a red ball* on *this table*»

«*Your last argument* is not accurate»

«Don't do *that*!»

«*The conference* was interesting»

Reference resolution

- Some terminology:
  - A linguistic expression used to perform reference is called a **referring expression**
  - The entity that is referred to is called the **referent**

Pierre

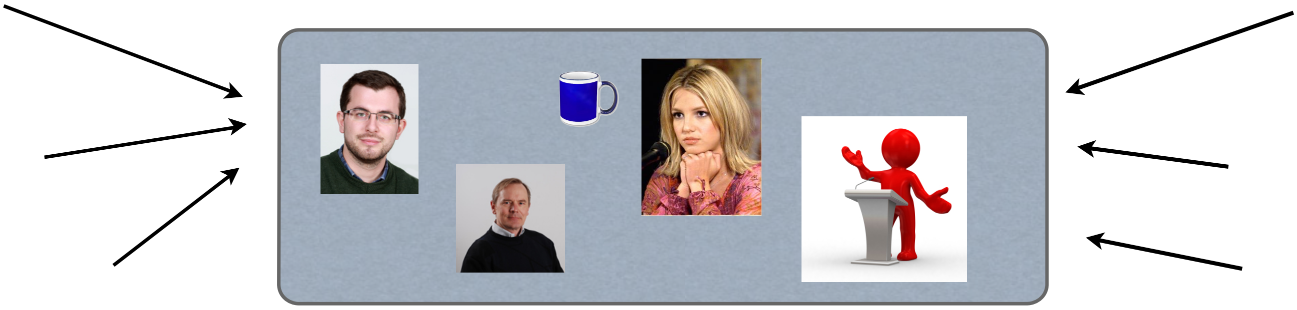
The INF5820 teacher

(referential ambiguity)

(coreference)

Reference resolution

- Reference resolution usually rely on a **discourse model** containing the set of *entities* that can be referred to
  - As well as their *relationships* with one another
  - The discourse model continuously change during the interaction (entities come and go, become more or less focused, etc.)
- In situated systems, the discourse model also contain objects or events in the shared environment



Reference resolution

- Types of referring expressions:

Indefinite noun phrases:	«*a beautiful goose*»
Definite noun phrases:	«*the conference*»
Pronouns:	«*she* gave a great talk»
Demonstratives:	«*this pen* is broken»
Names:	«*Jan Tore*»

- Choice of referring expression often depends on the *information status* of the referent:

in focus	>	activated	>	familiar	>	uniquely identifiable	>	referential	>	type identifiable
it		*that this, this N*		*that N*		*the N*		*indef. this N*		*a N*

[Gundel et al. (1993). «Cognitive status and the form of referring expressions in discourse», *Language*]

Reference resolution

- Various features can be used to resolve references:
  - Grammatical agreement (number, person, gender)
  - Saliency (recency of mention, visual salience, etc.)
  - Semantic constraints
- Based on these features and annotated training data, one can then train a *classifier*
  - *Binary* classification problem: given a referring expression A and a referent B the classifier determines whether A refers to B
  - Any supervised learning algorithm (e.g. log-linear models) will do

Outline

- Parsing spoken language
- Three challenges
- Reference resolution
- **Dialogue act recognition**
- Summary

Dialogue act recognition

- Dialogue acts:
    - «Functional units of a dialogue used by the speaker to change the context»
    - Extension of the concept of speech act to cover conversational phenomena (e.g. grounding)
    - Also called dialogue/conversational moves
- Various tagsets have been put forward

Dialogue act recognition

- DAMSL (Dialogue Act Markup in Several Layers) classifies dialogue acts in two dimensions:
    - **Forward-looking functions** are classical «speech acts», such as *assertions*, *directives*, *information request*, *commitments*, and *social conventions*
    - **Backward looking functions** «look back» at the previous utterances, and can signal *agreement*, *understanding*, or provide *answers*.

Dialogue act recognition

Assert	utt1	C	I need to travel in May.
Info-request, Ack	utt2	A	And, what day in May did you want to travel?
Assert, Answer	utt3 utt4	C	Ok uh I need to be there for a meeting that's from the 12th to the 15th.
Info-request, Ack	utt5	A	And you're flying into what city?
Assert, Answer	utt6	C	Seattle.
Info-request, Ack	utt7	A	And what time would you like to leave Pittsburgh?
Hold	utt8	C	Uh hmm I don't think there's many options for non-stop.
Accept, Ack	utt9	A	Right.
Assert	utt10		There's three non-stops today.
Info-request	utt11	C	What are they?
Assert, OO	utt12	A	The first one · · · . The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. · · ·
Accept, Ack	utt13	C	OK I'll take the second flight on the 11th.
Info-request, Ack	utt14	A	On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
Ack	utt15	C	OK.

Dialogue act recognition

- Again, one can train a classifier to recognise an utterance dialogue act based on a specific set of features:
  - **Lexical and syntactic features** (example: presence of «please» is a good indicator for a request)
  - **Prosody** (example: rising pitch in English is an indicator for a yes/no question)
  - **Dialogue structure** (example: «yeah» following a proposal is probably an agreement, while a «yeah» following an inform is most likely a backchannel)

Dialogue act recognition

- Search for most likely dialogue act *d** given the utterance *s*:

$$d^* = \underset{d}{\operatorname{argmax}}\, P(d|s) = \underset{d}{\operatorname{argmax}}\, \frac{P(s|d)P(d)}{P(s)}$$
$$= \underset{d}{\operatorname{argmax}}\, P(s|d)P(d)$$

- If we assume that the lexico-syntactic features *ls*, the prosody *p* and the previous dialogue act d_{t-1} are independent (Naive Bayes assumption), then:

$$d^* = \underset{d}{\operatorname{argmax}}\, (P(ls|d) \times P(p|d) \times P(d|d_{t-1}))$$

These 3 models can be directly estimated from annotated data

Outline

- Parsing spoken language
- Three challenges
- Reference resolution
- Dialogue act recognition
- **Summary**

Summary

- We have discussed today various topics related to *spoken language understanding:*
  - Parsing spoken dialogue can be done with shallow, grammar-based, or statistical methods
  - The presence of speech recognition errors, disfluencies and non-sentential utterances make this process more difficult than for text processing
  - But some pre-processing techniques can (partially) alleviate these problems

Summary

- We also covered reference resolution
  - The goal: find the most likely *referent* for a *referring expression*
  - Using features as agreement, salience, and semantic constraints, one can train a classifier to resolve referring expressions based on annotated data
- ... and dialogue act classification
  - Existence of various taxonomies of dialogue acts, some structured in several dimensions (e.g. DAMSL)
  - One can also train a classifier to recognise dialogue acts based on lexico-syntactic, prosodic and dialogue features

Next Wednesday

- Next Wednesday, we'll talk about **dialogue management**
  - How do we decide what is the best thing to do/say at a given point in the interaction?
  - What are the different ways to describe this decision-making process?
  - Can we learn «the best thing to do» based on training data (supervised learning) or the system's own experience (reinforcement learning)?