Spoken Language Understanding strategies developed at the University of Avignon: For a better integration of ASR and SLU processes

Frédéric Béchet

LIA, Université d'Avignon

Introduction

- Spoken Language Understanding ?
  - Everything going beyond word transcriptions
    - Structure, theme, entities, etc.
  - Corpus-based method = Need for observations
    - Direct observations
      - Linked to an action of the speaker
    - Indirect observations
      - Manual annotations of spoken message

SLU vs. Text processing

- SLU = ASR + text processing ?
  - Text documents vs. Speech utterances
  - Automatic transcripts
    - ASR issues
      - Uncertainty, misrecognition, unknown words
    - Partial information
      - All prosodic information missing
    - No structure = stream of words
  - Text
    - “finite” object
    - Text + structure + “graphical” information

SLU vs. Text processing

- Main issues
  - Text
    - “open world”
    - Capacity of handling new phenomenon
      - Words, compounds, entities
    - Need: Generalization capabilities of the models
  - ASR transcript
    - “closed world”
    - ASR lexicon+Language Model define this “world”
    - No unknown words (just misrecognitions !!)

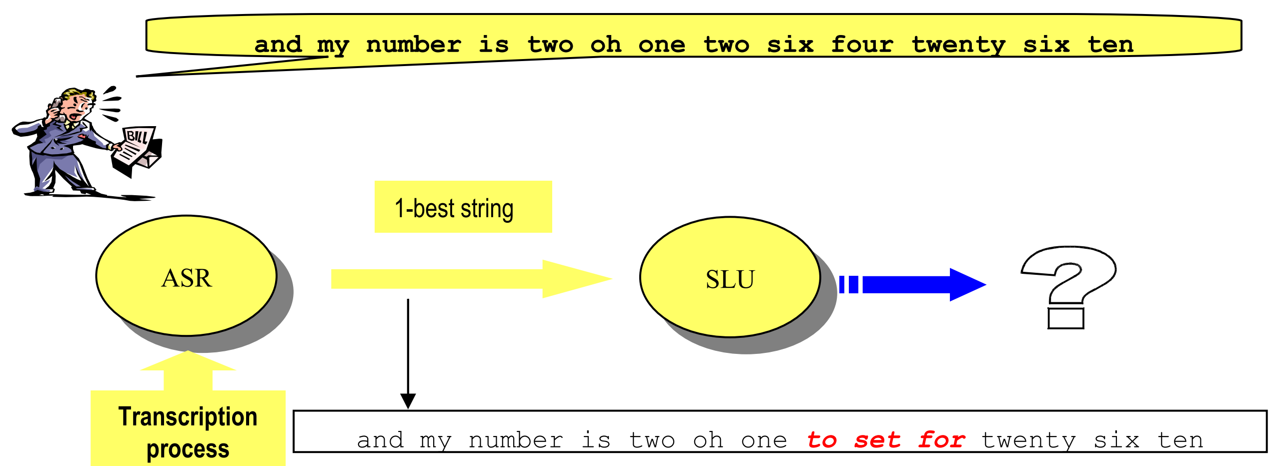
      => no generalization needed
    - Need: robust detection of the expected information
      - Confidence estimation

SLU strategies

- 3 modules
  - ASR
    - From speech to words
  - SLU
    - From speech+words to interpretations
  - "Manager"
    - To exploit the interpretations
      - Dialog manager, speech mining, etc.
- Need for contextual information
  - To identify what is expected
  - At each level of the process: ASR, SLU, Manager
    - To rescore hypotheses, for the decision process

SLU strategies: two main approaches

- « sequential approach »
  - ASR => SLU => Manager
    - ASR module produces a text document
    - SLU module processes this text document
    - Manager = exploits SLU output



SLU strategies: two main approaches

- « integrated approach »
  - ASR ⇔ SLU ⇔ Manager
  - All 3 processes should collaborate
    - Definition of a context
    - ASR+SLU+Manager: tuning according to the context
    - ASR output = multiple hypothesis (word lattice)
    - SLU = from a word lattice to an « interpretation lattice »
    - Manager = decision strategy on multiple hypothesis output

Applications, corpus ?

- « artificial corpus »
  - Collected through evaluation program (Ex: ATIS, MEDIA)
  - Manual annotations
  - Limited size
  - Application domain
    - Spoken dialogue systems, question answering, speech doc. retrieval
- « real life corpus »
  - Collected from real users of a speech-service
    - Ex: AT&T How May I Help You?, France Telecom Voice Services
  - Annotations = automatic/manual/none
  - Unlimited size
  - Application domain
    - Call-centers, Audio messages, Deployed SDS

Applications, corpus ?

- Main differences
  - Artificial corpus
    - controlled conditions
    - cooperative speakers
    - => little “out-of-domain” data
  - Real life corpus = real life issues !!
    - Very spontaneous speech
    - Very large variability
      - Speech: accents, language
      - Usage: different classes of users (new and regulars)
    - Unpredictable behaviors
      - Comments, incoherence

Context of this study

- Collaboration with France Telecom R&D
  - SLU for FT 3000 voice service
  - Speech mining
    - Spoken survey of customers opinions
- French program Technolangue/Evalda/Media
  - Concept decoding (Spoken dialog systems)
  - Reference resolution
- European Project STREP LUNA
  - Integrated approach for SLU
  - Semantic composition

LUNA

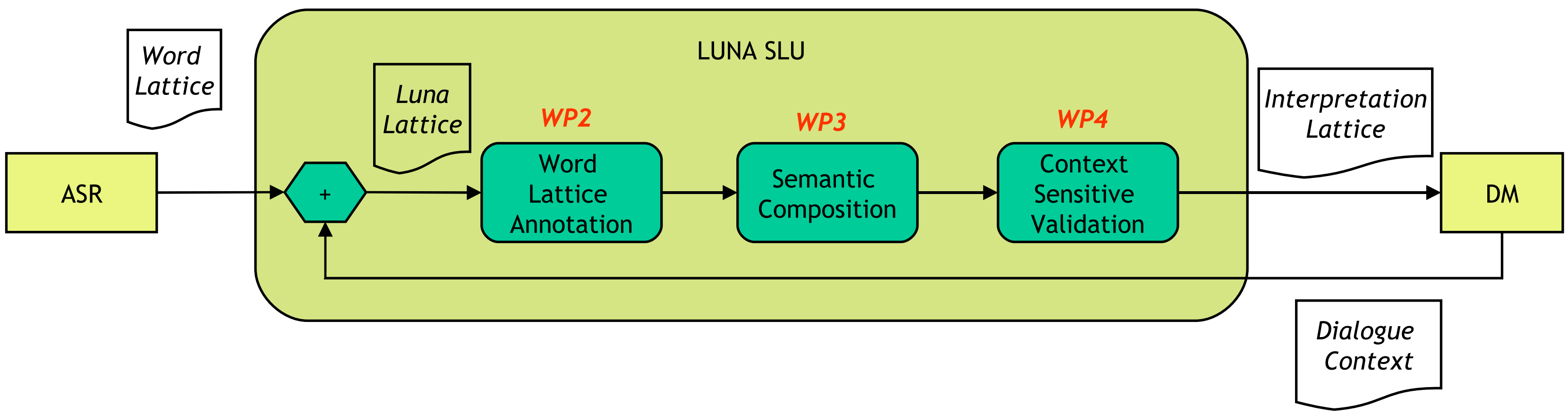
- FP6 European project: LUNA
  - spoken Language UNderstanding in multilinguAl communication systems
  - September 2006
- Goal
  - Build robust multilingual SLU strategies
  - Five main objectives
    - Language Modelling for Speech Understanding;
    - Semantic Modelling for Speech Understanding;
    - Automatic Learning (including Active and On-Line Learning);
    - Robustness issues for SLU;
    - Multilingual portability of SLU components.
- Partners
  - Loquendo, RWTH Aachen, University of Trento, University of Avignon, France Telecom R&D, CSI-Piemonte, Polish-Japanese Institute of Information Technology, Institute of Computer Science - Polish Academy of Sciences

SLU models in LUNA

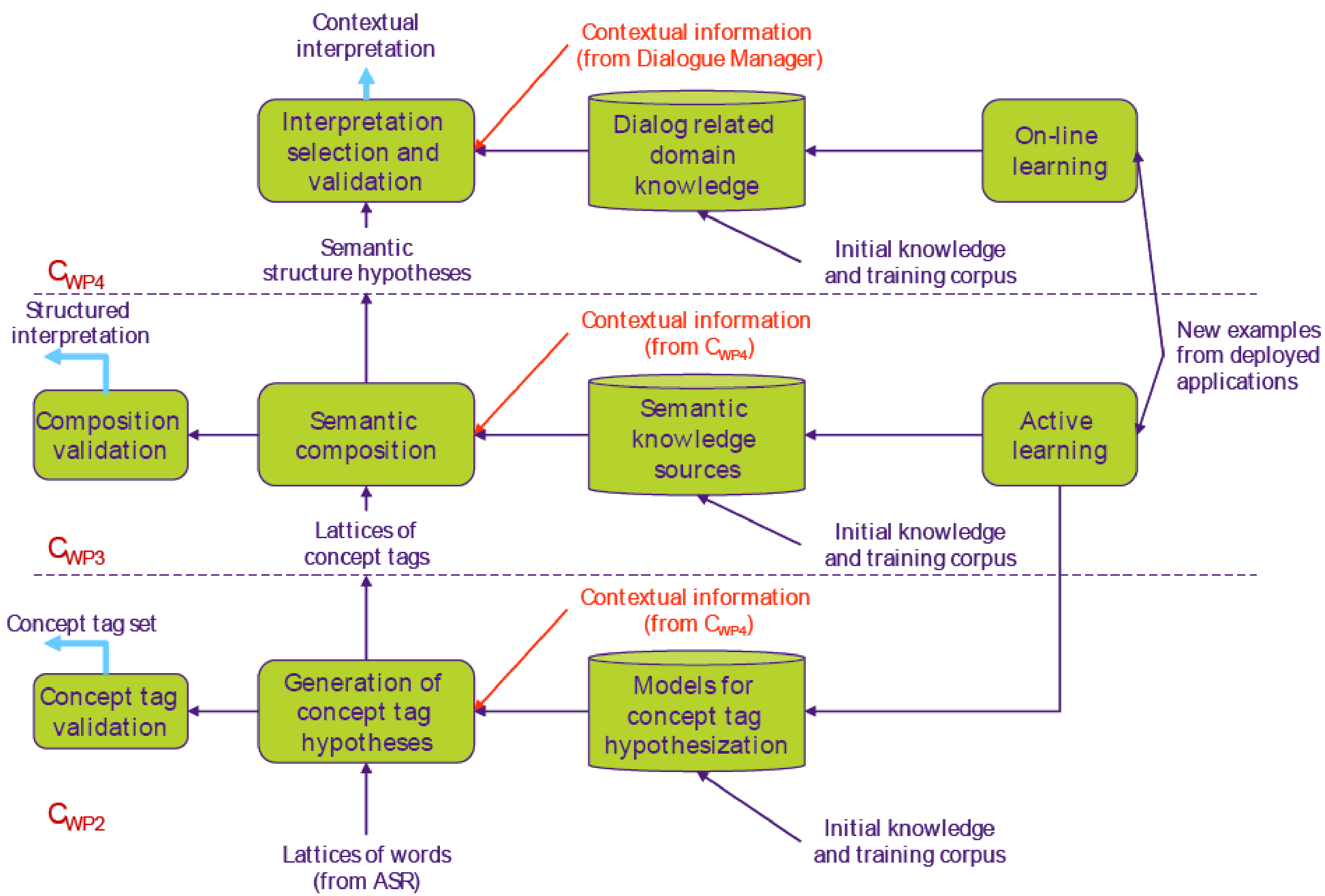
- Multi level semantic representation
  - Concept decoding: from words to concepts
  - Semantic composition: from concepts to interpretations
  - Coreference / Anaphoric relation resolution
  - Speech acts
- Corpus annotation on these levels
  - Concepts
    - word+POS tag+chunk+ Ontology in OWL
  - Interpretations
    - Framenet-like approach
  - Reference resolution
    - ARRAU framework
  - Speech acts
    - Subset of DAMSL

LUNA: an integrated approach

– Process
  - From a word lattice to an entity lattice
  - From an entity lattice to an interpretation lattice
  - With references, with speech acts
  - Each level using contextual information
    – A priori information on the application context
    – Dynamic information provided bt the dialog manager
– Corpus based + knowledge based methods



LUNA architecture



First level: words to “concepts”

- concepts=entities, attribute-value, …
- Translation from words to concepts
  - « traditional » task for NLP on text (shallow parsing)
  - Particularities on speech messages
    - text = open world => need for generalization
    - ASR transcriptions = closed world, “no” OOV words
- Strategies
  - Leaves in a parse tree
  - Hand-written rules
  - Translation model (statistical translations)
  - Tagging model
    - HMM, Conditional Random Field, Dynamic Bayesian Network
  - Classification task
    - Boosting, MaxEnt, SVM, etc.

First level: words to “concepts”

- Processing speech utterance
  - Integrated search
    - Best sequence of words / of concepts
    - Constraining the transcription with concept information
    - From a word lattice to a concept lattice
  - Integrating contextual information
    - What is expected?
      - Local context
      - Global context

Example (global context)

I wanna know why I was charged on **September sixth 11 dollars 63 cents** for calling **8 5 6 2 1 6 5 5 2 1 Clementon New Jersey** for **1 minute**

PHONE BILL SEPTEMBER 2001

DATE	PHONE#	DURATION	PLACE	AMOUNT
09062001	8562165521	01:00	Clementon, NJ	11.63
….	….	….	….	….
….	….	….	….	….

Exemple: AT&T How May I Help You? tm

Example (local context)

```
system>  in Marseille I propose the Hotel la Fanette
         and the Hotel du Port
user>    where is the Hotel la Fanette?
ASR>     where is the Hotel Lafayette
```

First level: words to “concepts” : strategy

- Integrated search
  - “concept” model as a Language Model for ASR
  - HMM Tagger for dealing with ambiguities on the hypotheses obtained
- Integrating contextual information
  - Global context
    - Modeling all the “expected” concepts (ASR lexicon)
    - From corpus analysis + a-priori knowledge
  - Local context
    - Conditional probabilities on the concepts, cache-based models
    - Integrating dialog states in the model
- Output
  - Lattice of concepts
  - Structured list of hypotheses
- Discriminant classification process
  - Classifiers, CRF

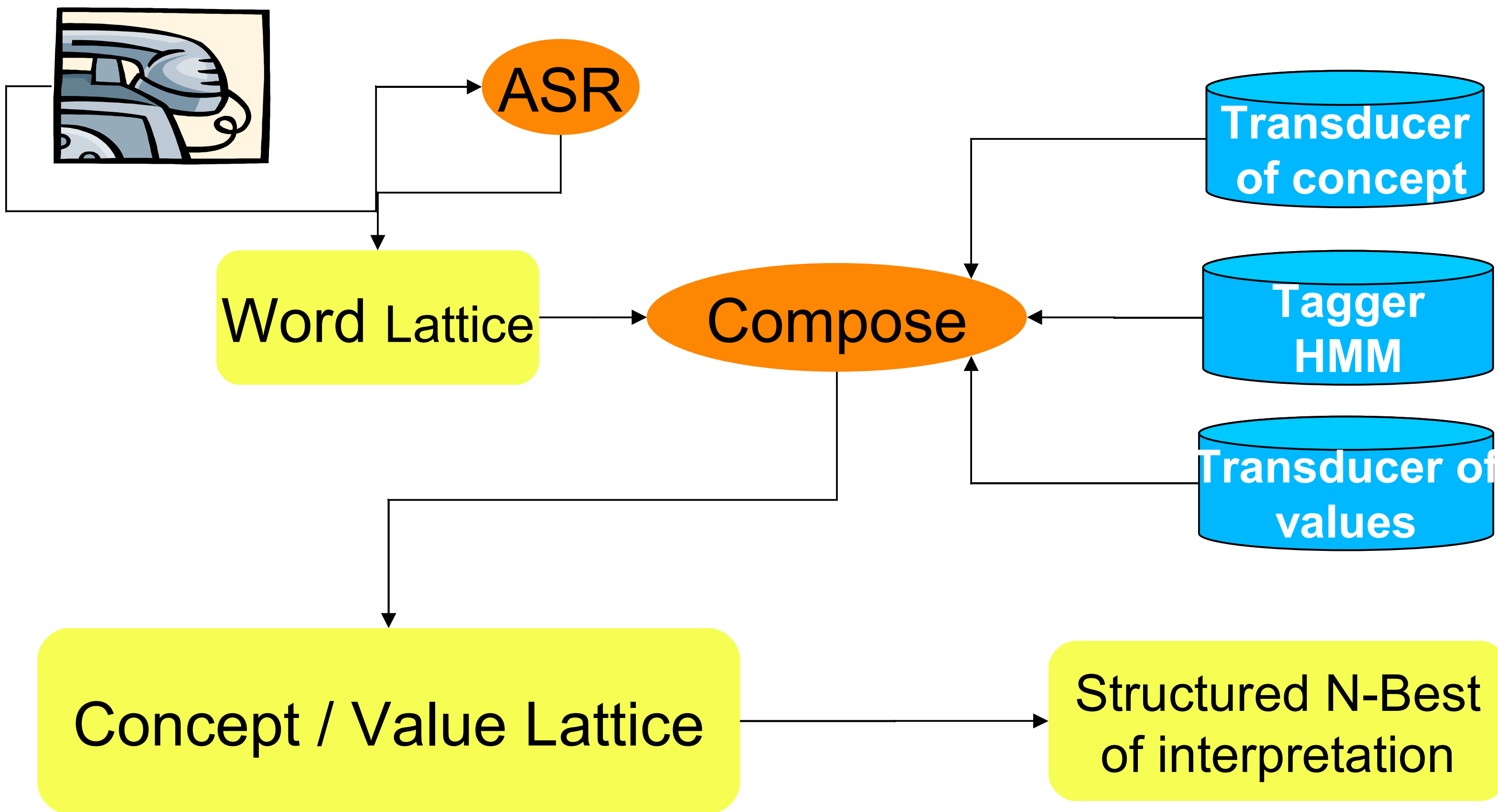
Application: the MEDIA spoken dialog corpus

- Tourism info + hotel booking services
- French Technolangue Project
- Manual annotations
  - word + concept transcriptions
- Corpus
  - Wizard of Oz
  - 250 speakers, 5 dialogues each
  - 1250 dialogs

Example

N	W	C	value
0	uh	null	
1	yes	answer	yes
2	the	RefLink	singular
3	hotel	BDObject	hotel
4	which	null	
5	price	object	payment-amount
6	is below	comparative-payment	below
7	hundred and ten	payment-amount-int	110
8	euros	payment-currency	euro

Strategy


ASR
Word Lattice
Compose
Transducer of concept
Tagger HMM
Transducer of values
Concept / Value Lattice
Structured N-Best of interpretation

Example of structured n-best list

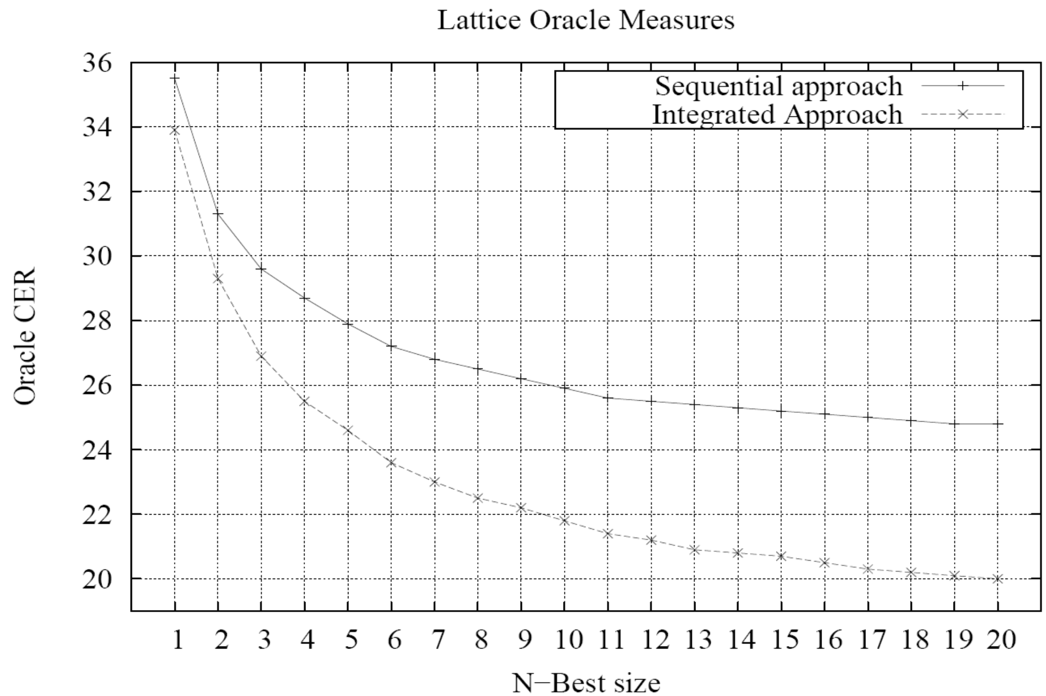
"je voudrais réserver à l'hôtel de Genève à Paris pour le 18 Septembre"

Int1*	Command-Task	Name-Hotel	Localisation-City	Time-Date
Values 1	Reservation	Geneve	Paris	18/09
Values 2	Reservation	Unknown	Geneve	18/09
Int2*	**Command-Task**	**ObjectDB**	**Localisation-City**	**Time-Date**
Values 1	Reservation	Hotel	Geneve	18/09
Values 2	Reservation	Hotel	Paris	18/09

Evaluation

- Test corpus: 200 dialogues
- Concept tagset: 83 concept tags
- Measures: Word Error Rate (WER) + Concept Error Rate (CER)+Oracle CER
- 2 strategies: Sequential approach (Seq) / Integrated approach (Int)

Score	CER %		WER %	
	Seq	Int	Seq	Int
Lattice ASR	35,5	33,7	33,5	33,4
Trans.	20,5	20,5	0	0



Second level: "concepts" to "interpretations"

- Semantic composition
  - Logical rules applied on the concepts
  - Composition of "basic concepts" into structured entities
    - ex: LUNA FrameNet-like predicate structure
  - Input
    - N-best lists of concept strings
    - Concept lattice
      - Rules encoded as FSM
- Coreference / Anaphoric relation resolution
  - Tagging + rule based approach
- Speech acts
  - Classification task

FT 3000 Voice Agency service

- Service
  - obtain information about FT services
    - purchase almost 30 different services
  - access account
    - check consumption, pay bills
    - call forwarding
    - voice messaging
- Deployed since October 2005
- Corpus collected daily

FT 3000 Voice Agency service

- Semantic model
  - Verbateam SLU system
  - 2-level model
    - 1st level: word to concept
      - Concept = sequence of keywords representing services
      - ~100 concepts. Ex:
        - **illimités dix numéros : [I10N]**
        - **trente_et_un dix : [AtoutPartout]**
      - Concept = local grammars representing a request
      - ~300 grammars. Ex:
        - **au fur et à mesure : [Rapidement]**
        - **comment diminuer : [Limiter]**

FT 3000 Voice Agency service

- 2nd level: concept to interpretation
  - Logical rules on the concepts
  - Ordered list: first match
  - ~3000 rules
  - Example:

```
((Resilier|Annuler|Supprimer|Arreter|Plu)

# ((Appel|Appelle|Telephone|Telephoner) & Frequent &
Domicile))

=> {Gest(Resilier,Ambi(AtoutsPlus,HeureLocale,ForfaitLocal))}
```

From a sequential to an integrated SLU

- Deployed system
  - Sequential, non stochastic SLU
- Integrated SLU trained on the automatic annotations
  - ASR output = word lattice
  - Concepts = local grammars = FSM (AT&T FSM Library)
  - Concept tagger = HMM-based tagger
    - Encoded as a FSM Language Model (AT&T GRM Library)
  - Interpretation rules
    - Encoded as transducers
      - Concept tags as input
      - Rule ID + rank in the rule database
  - Dialog states
    - Language model on the dialog states
      - Encoded as an FSM

Stochastic Model

$S = \{S_0, S_1, \ldots, S_k\}$ Sequence of dialog states
$Y = \{Y_1, Y_2, \ldots, Y_k\}$ Sequence of utterances
$\Gamma = \{\Gamma_1, \Gamma_2, \ldots, \Gamma_k\}$ Sequence of interpretations
$C = c_1, c_2, \ldots, c_n$ Basic concept string
$W = w_1, w_2, \ldots, w_l$ Word string

$$P(S|Y) = \sum_{\Gamma} P(S\Gamma|Y)$$

$$P(S_k\Gamma_k|S_{1,k-1}\Gamma_{1,k-1}Y_k) \approx \quad P(S_k|\Gamma_k S_{k-1})P(\Gamma_k|Y_k)$$

Stochastic Model

$$P(\Gamma_k|Y_k) = \sum_{C_k,W_k} P(\Gamma_k C_k W_k|Y_k)$$

$$\approx \sum_{C_k,W_k} P(\Gamma_k|C_k)P(Y_k|W_k)P(W_k)^{\alpha-\beta}P(C_k,W_k)^{\beta}$$

$P(S_k|\Gamma_k S_{k-1}) \longrightarrow$ Bigram Language model on the dialog states = ***D***

$P(\Gamma_k|C_k) \longrightarrow$ Composition rules: 0 / 1 = ***R***

$P(Y_k|W_k) \longrightarrow$ Acoustic Model = ***A***

$P(W_k) \longrightarrow$ Trigram word Language Model = ***W***

$P(C_k,W_k) \longrightarrow$ word, concepts tagger = ***C***

Implementation

- With
  - Transducer interpretation+context => dialog state = ***S***
  - Bigram Language model on the dialog states = ***D***
  - Composition rules: 0 / 1 = ***R***
  - Language Model on the word+concept = ***C***
  - Trigram word Language Model = ***W***
  - Word-to-Concept transducer = ***T***
  - Word lattice from ASR = ***L***

$$\hat{I} = \mathrm{bestpath}\left([(L_1 \circ W \circ T \circ C \circ R \circ S) \ldots (L_n \circ W \circ T \circ C \circ R \circ S)] \circ D \right)$$

$\hat{I}$: best interpretation at turn *n*

Processing « real » corpora

- Dealing with different kind of speech
  - Speech/non speech
  - Speech out-of-domain/speech in domain
  - Speech with a valid content/invalid content
- Evaluation ?
  - the performance of the service
    - Difficult in batch mode
  - each module separately
    - Which impact on the global performance?
  - On what kind of speech?
    - Every signal segment detected
    - Only on the meaningful segments

Processing « real » corpora

- Strategy proposed
  - ASR: Multiple processes, multiple outputs
    - 1best, word lattice, confusion network
  - Detecting as soon as possible non relevant segment
  - Applying « sophisticated » SLU only on reliable segments
- Main feature
  - 1st pass LM detecting in-domain/out-of-domain speech
  - Confidence measures from the confusion network
  - Detection of « reliable » segments
  - Structured n-best list of hypothsis on these segments
  - Possible queries from the manager

Detection Out-of-Domain segments

- Modeling out-of-domain?
  - *Comments* from the callers. Ex:
    - “can you close the door please”
    - “what am I suppose to say now”
    - “I can’t believe it”
    - “you **** ****”
- Specific 2-level language model
  - 1 general LM + 1 LM trained on the comment segments
  - Ex: **`<s> w1 <comment> w2 w3 </comment> w4 </s>`**

$$P^{G+OOD}(w_1, w_2, w_3, w_4) = P^G(w_1|start) \times P^G(\_OOD\_|w_1) \times P^{OOD}(w_2|start) \times P^{OOD}(w_3|w_2) \times P^{OOD}(end|w_3) \times P^G(w_4|\_OOD\_)$$

Experiment 1

- Corpus
  - Training
    - 44K utterances for LM (word and concept)
    - 7.4K dialogues (dialog state LM)
  - Test
    - 816 dialogues / 1950 utterances
- User profiles
  - Register users
    - 80% of the calls, 60% of the utterances
  - New users
    - Longer dialogs, more comments

Experiment 1

- User profiles: experienced vs. new users

	other	transit
# dialogues	350	467
# utterances	1288	717
# words	4141	1454
av. dialogue length	3.7	1.5
av. utterance length	3.2	2.0
OOV rate (%)	3.6	1.9
disfluency rate (%)	2.8	2.1

	other	transit
# dialogues	350	467
# utterances	1288	717
# OOD comments	137	24
OOD rate (%)	10.6	3.3
dialogues with OOD (%)	14.3	3.6

Experienced users prefer keywords and don't make comments !!

Experiment 1

- Results
  - OOD LM is very useful on the *other* dialogues
  - Small gain in IER with integrated approach

IER	all	other	transit
size	1953	734	1219
LM^{G}	16.5	22.3	13.0
LM^{G+OOD}	15.0	18.6	12.8

corpus	**all**		
error	*WER*	*CER*	*IER*
strat1	40.1	24.4	15.0
strat2	38.2	22.5	14.5
strat3	38.3	22.5	14.7

corpus	**other**		
error	*WER*	*CER*	*IER*
strat1	48.8	34.7	18.6
strat2	47.6	34.2	18.9
strat3	47.9	34.4	19.4

corpus	**transit**		
error	*WER*	*CER*	*IER*
strat1	31.8	18.2	12.8
strat2	29.3	14.2	11.8
strat3	29.1	14.0	11.8

Experiment 1

- Using multiple hypotheses output

IER	all		other		transit	
consensus	IER	cover	IER	cover	IER	cover
1	15.0	100%	18.6	100%	12.8	100%
1∧2	12.7	88.7%	15.1	86.4%	8.7	92.8%
1∧2∧3	12.0	87.6%	14.3	84.9%	8.3	92.3%

- Can be used to detect problematic dialogues

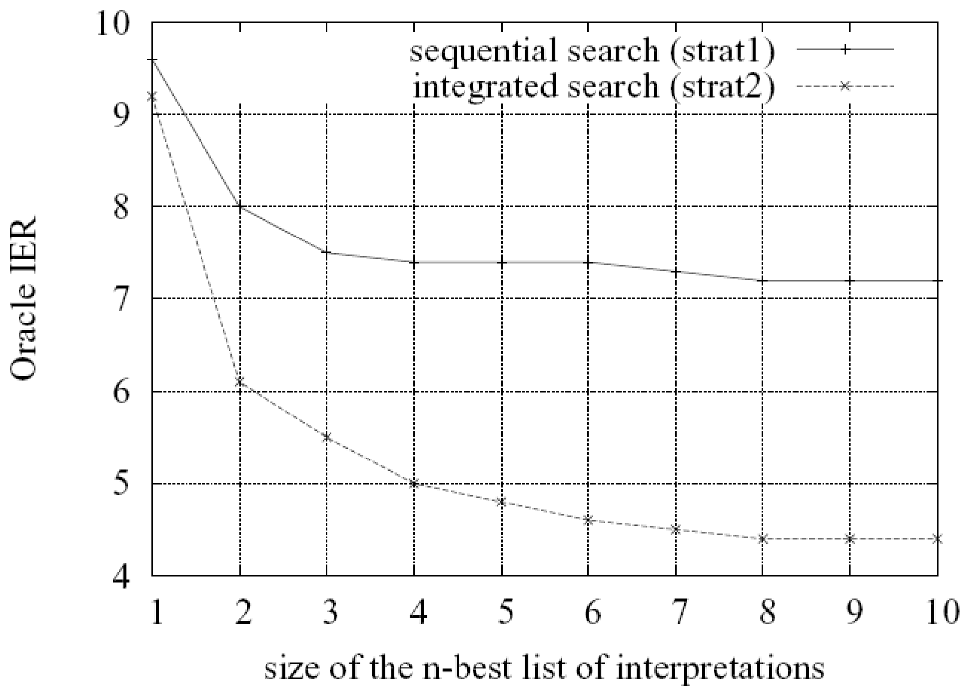
Experiment 1

- Oracle

level	*1-best*	*Oracle hyp.*
WER	33.7	20.0
CER	21.2	9.7
IER	13.0	4.4

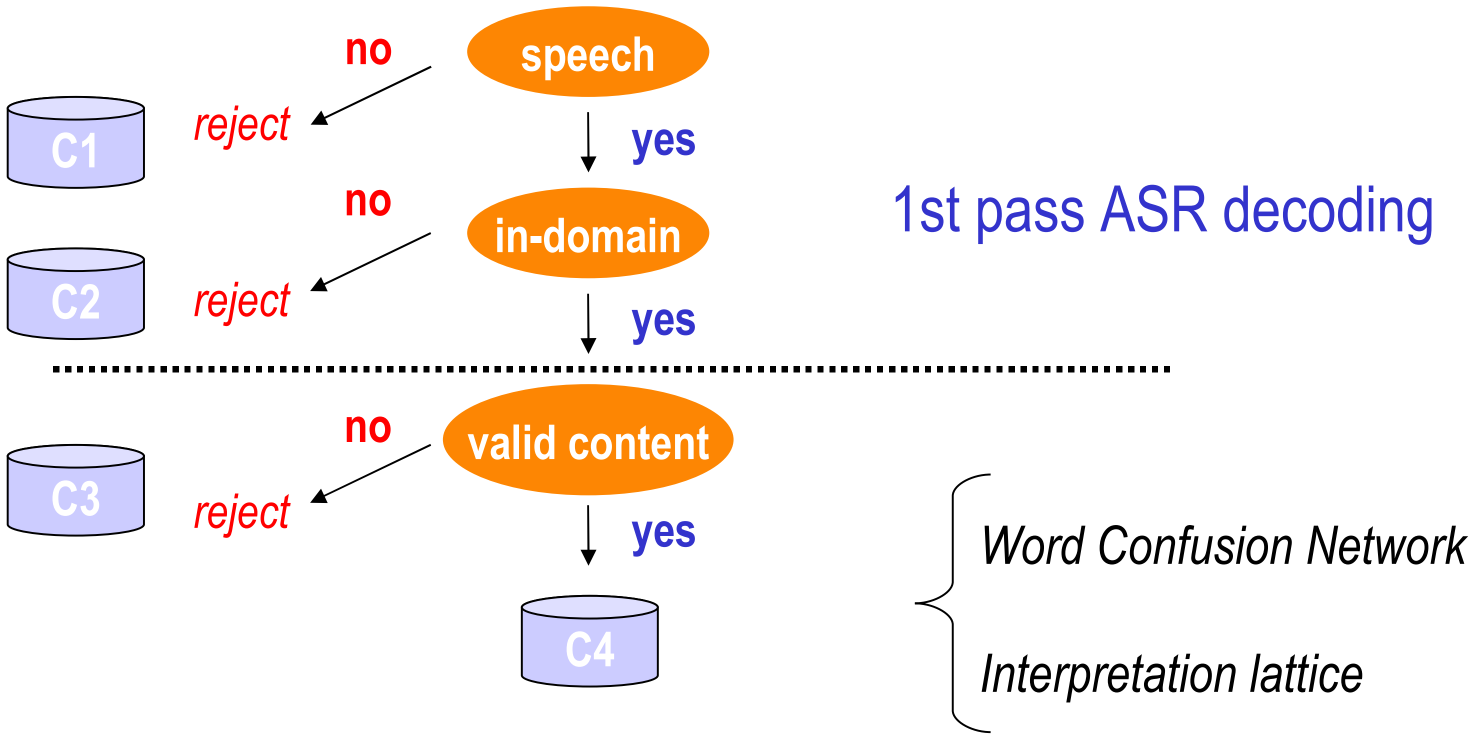
- sequential *vs* integrated oracle error rates

	IER
from word Oracle	9.8
from concept Oracle	7.5
interpretation Oracle	4.4



Experiment 2

- Detecting as soon as possible «empty» utterances
- Using «rich» search space only on reliable segments



Experiment 2

Category	# utterances
C1: Non-Speech detections	1333
C2: Out-of-Domain speech	674
C3: In-Domain speech without interpretation	355
C4: In-Domain speech with interpretation	4139
Total	6501

Test corpus: 3200 dialogs, 6500 utterances

		WL 1-best	CN 1-best	WL Decoding	CN Decoding
False acceptance	*FA* on C1	6.5 %	**2.6 %**	22.8 %	20.1 %
	FA on C2	7.8 %	**5.3 %**	13.0 %	13.7 %
	FA on C3	2.9 %	**2.3 %**	6.3 %	7.0 %
Interpretation Error Rate	*Sub+FR* on C4	8.7 %	10.6 %	**6.5 %**	8.6 %

Experiment 2

total	Baseline (1-best)	*Strat1* (CN)	*Strat2* (WL)
FA	17.2 %	8.8 %	8.8 %
Sub	6.1 %	5.6 %	4.1 %
FR	2.7 %	5.2 %	5.2 %
IER	**26.0 %**	**19.6 %**	**18.1 %**

Strat1 : sequential approach, rejection on the 1-best
Strat2 : rejection on the consensus hyp. + SLU in the WCN
Strat3 : rejection on the consensus hyp. + SLU in the WL

Conclusions

- For a better integration of the upstream and downstream processes
- « context » must be used at each level of the SLU processes
- Confidence measures and rejection strategies are crucial for processing «realistic» utterances
- Multiple hypotheses strategies involving discriminant approaches