A New SLT Decoder based on Confusion Networks

Nicola Bertoldi and Marcello Federico

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo (Trento), Italy

{bertoldi,federico}@itc.it

Outline

Outline

- **Spoken Language Translation**

Outline

- Spoken Language Translation
- Approaches

Outline

- Spoken Language Translation
- Approaches
- Confusion Network (CN)

Outline

- Spoken Language Translation
- Approaches
- Confusion Network (CN)
- CN-based Translation Model

Outline

- Spoken Language Translation
- Approaches
- Confusion Network (CN)
- CN-based Translation Model
- CN-based Decoder

Outline

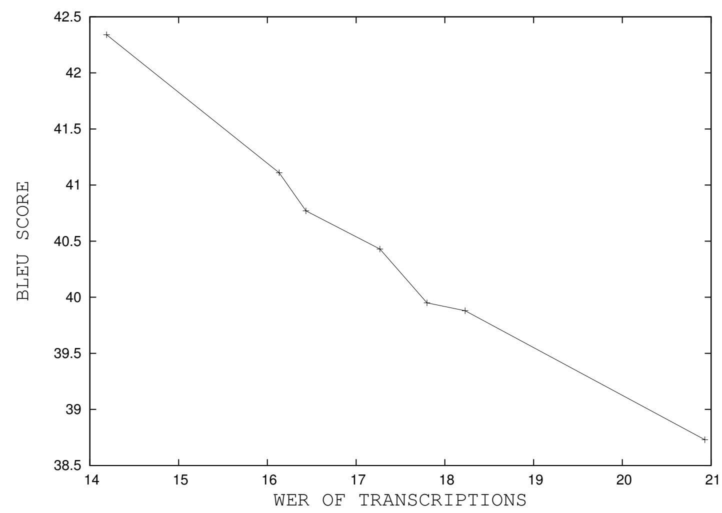
- Spoken Language Translation
- Approaches
- Confusion Network (CN)
- CN-based Translation Model
- CN-based Decoder
- Evaluation

Spoken Language Translation

Spoken Language Translation

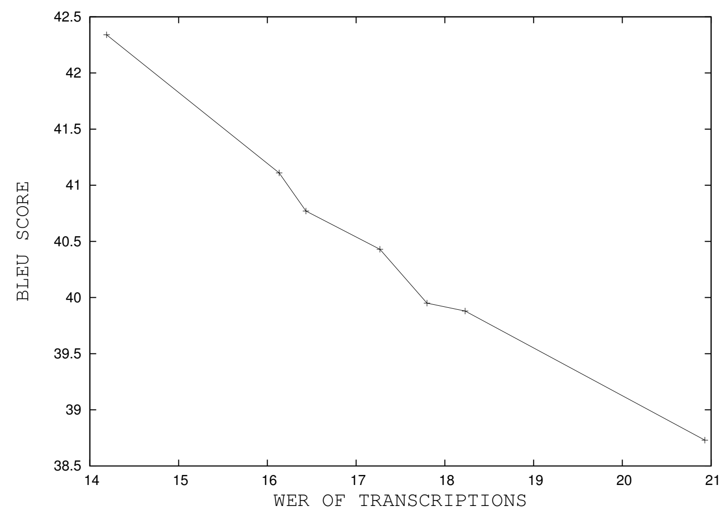
- Translation of speech input
  - spontaneous speech phenomena:
    repetitions, hesitations
  - recognition errors:
    syntax, meaning

Spoken Language Translation



- **Translation of speech input**
  - **spontaneous speech phenomena: repetitions, hesitations**
  - **recognition errors: syntax, meaning**
- **Automatic Speech Recognition and Machine Translation**
  - **strong correlation between recognition and translation quality**
  - **ASR WER decreases in a set of hypotheses**

Spoken Language Translation



- **Translation of speech input**
  - **spontaneous speech phenomena: repetitions, hesitations**
  - **recognition errors: syntax, meaning**
- **Automatic Speech Recognition and Machine Translation**
  - **strong correlation between recognition and translation quality**
  - **ASR WER decreases in a set of hypotheses**
  - **idea: exploitation of more transcriptions**

Statistical Spoken Language Translation

Statistical Spoken Language Translation

Given a speech input o in the source language, and the set $\mathcal{F}(\mathrm{o})$ of its possible transcriptions, find the best translation through the following approximate criterion:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \Pr(\mathbf{e} \mid \mathbf{o}) \approx \arg\max_{\mathbf{e}} \max_{\mathbf{f} \in \mathcal{F}(\mathbf{o})} \Pr(\mathbf{e}, \mathbf{f} \mid \mathbf{o})$$

Statistical Spoken Language Translation

Given a speech input $\mathbf{o}$ in the source language, and the set $\mathcal{F}(\mathbf{o})$ of its possible transcriptions, find the best translation through the following approximate criterion:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \Pr(\mathbf{e} \mid \mathbf{o}) \approx \arg\max_{\mathbf{e}} \max_{\mathbf{f} \in \mathcal{F}(\mathbf{o})} \Pr(\mathbf{e}, \mathbf{f} \mid \mathbf{o})$$

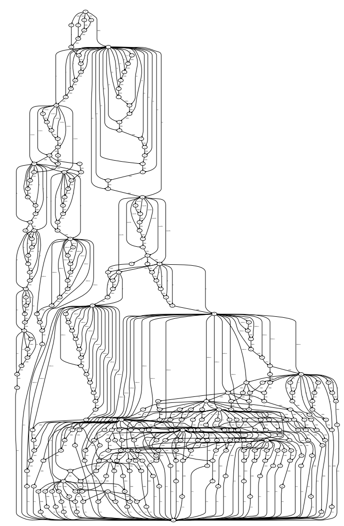
- $\Pr(\mathbf{e}, \mathbf{f} \mid \mathbf{o})$ **speech translation model**
  - **acoustic and translation features**

Statistical Spoken Language Translation

Given a speech input o in the source language, and the set $\mathcal{F}(\mathrm{o})$ of its possible transcriptions, find the best translation through the following approximate criterion:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \Pr(\mathbf{e} \mid \mathbf{o}) \approx \arg\max_{\mathbf{e}} \max_{\mathbf{f} \in \mathcal{F}(\mathbf{o})} \Pr(\mathbf{e}, \mathbf{f} \mid \mathbf{o})$$

- $\Pr(\mathrm{e}, \mathbf{f} \mid \mathrm{o})$ **speech translation model**
  - **acoustic and translation features**
- $\mathcal{F}(\mathrm{o})$ **is an ASR word graph (WG):**
  - **huge amount of transcription hypotheses**
  - **complex structure**



Approaches

Approaches

- **1-best Decoder: a text MT system only translates the best transcription of the ASR. No use of multiple transcriptions.**

Approaches

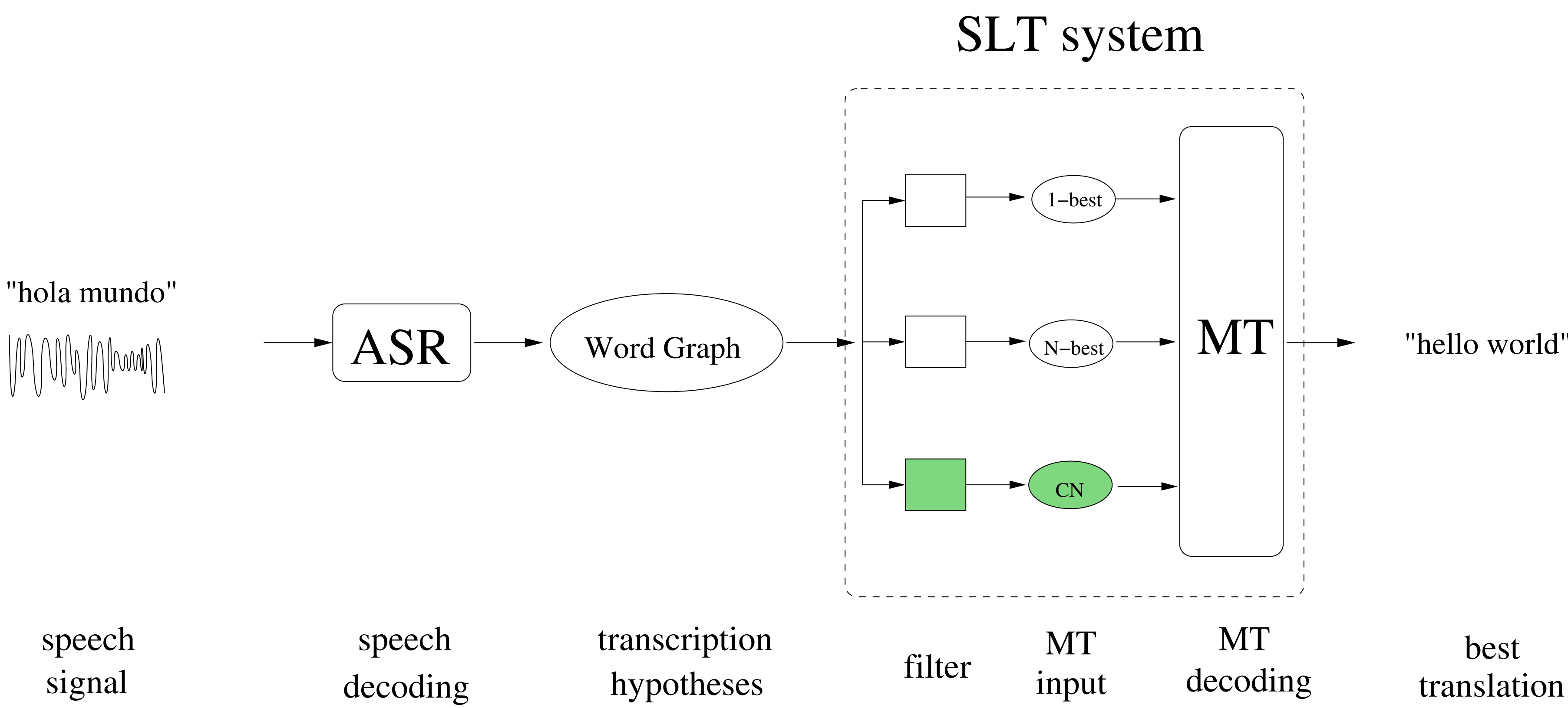
- **1-best Decoder: a text MT system only translates the best transcription of the ASR. No use of multiple transcriptions.**
- **N-best Decoder: N hypotheses are translated by a text MT decoder and reranked according to ASR scores, e.g. of acoustic and source language models. It does not advantage from overlaps among N-best.**

Approaches

- **1-best Decoder: a text MT system only translates the best transcription of the ASR. No use of multiple transcriptions.**
- **N-best Decoder: N hypotheses are translated by a text MT decoder and reranked according to ASR scores, e.g. of acoustic and source language models. It does not advantage from overlaps among N-best.**
- **Finite State Transducer: both ASR and MT models are merged into one finite-state network and a transducer decodes the input speech signal in one shot. Difficult scaling up to large domains.**

Approaches

- **1-best Decoder: a text MT system only translates the best transcription of the ASR. No use of multiple transcriptions.**
- **N-best Decoder: N hypotheses are translated by a text MT decoder and reranked according to ASR scores, e.g. of acoustic and source language models. It does not advantage from overlaps among N-best.**
- **Finite State Transducer: both ASR and MT models are merged into one finite-state network and a transducer decodes the input speech signal in one shot. Difficult scaling up to large domains.**
- **Confusion Network Decoder: an approximate WG is extracted from the ASR output and is directly translated. It exploits overlaps among hypotheses.**

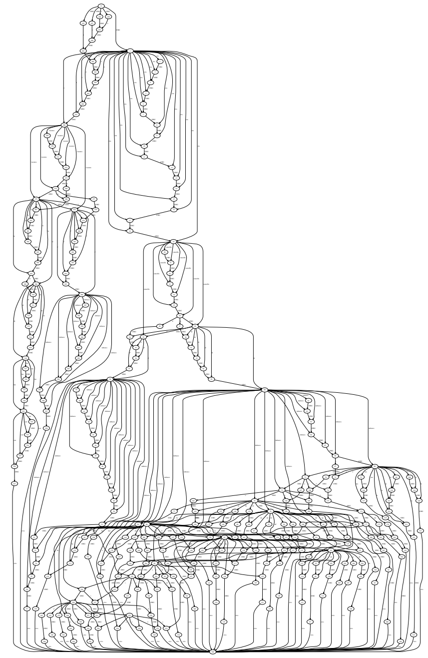

SLT system
"hola mundo"
ASR
Word Graph
1−best
N−best
CN
MT
"hello world"
speech signal
speech decoding
transcription hypotheses
filter
MT input
MT decoding
best translation

Confusion Network

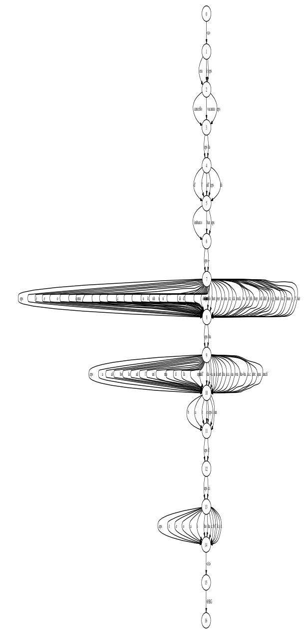
Confusion Network

- **A Confusion Network (CN) approximates a WG by shrinking into a unifilar WG (Mangu 1999)**

Word Graph



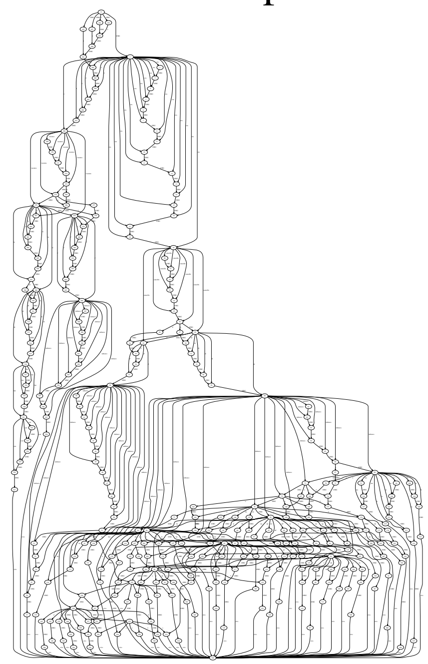
Confusion Network



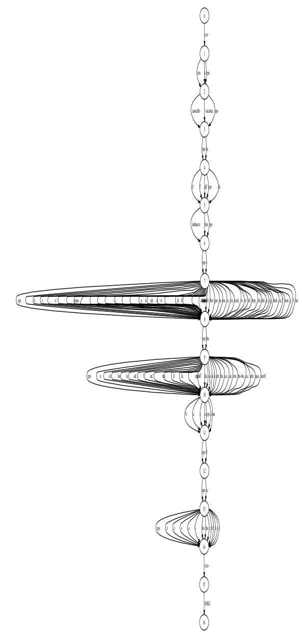
Confusion Network

- A Confusion Network (CN) approximates a WG by shrinking into a unifilar WG (Mangu 1999)
- Representation through a compact table

Word Graph



Confusion Network

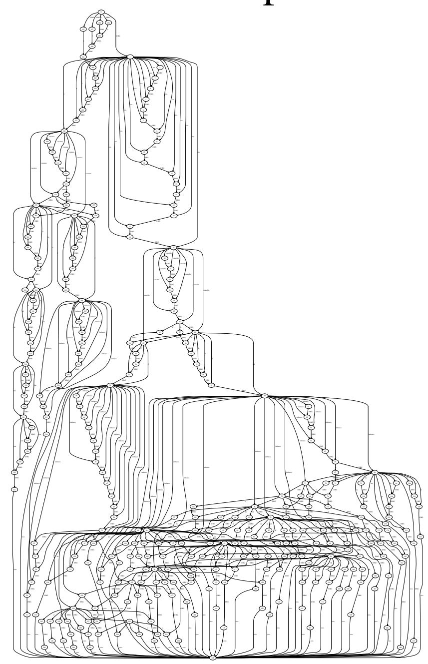


era 0.997	**cancello** 0.995	ϵ 0.999	**di** 0.615	**imbarco** 0.999	...
è 0.002	**vacanza** 0.004	**la** 0.001	**d'** 0.376	**bar** 0.001	
ϵ 0.001	ϵ 0.002		**l'** 0.002		
			...		
			ϵ 0.001		

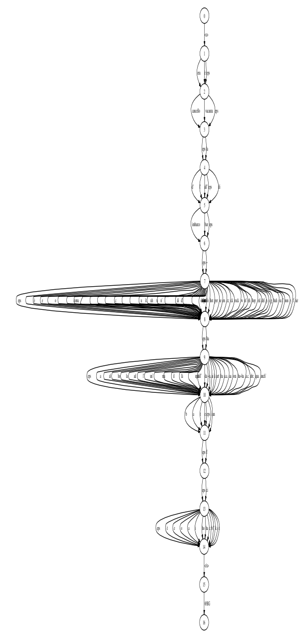
Confusion Network

- A Confusion Network (CN) approximates a WG by shrinking into a unifilar WG (Mangu 1999)
- Representation through a compact table
- Each path corresponds to a hypothesis
- CN contains more paths than WG
- Possible insertion of ϵ words
- Posterior probs for single words
- Likelihood for each hypothesis

Word Graph



Confusion Network



era 0.997	**cancello** 0.995	ϵ 0.999	**di** 0.615	**imbarco** 0.999	...
è 0.002	**vacanza** 0.004	**la** 0.001	**d'** 0.376	**bar** 0.001	
ϵ 0.001	ϵ 0.002		**l'** 0.002		
			...		
			ϵ 0.001		

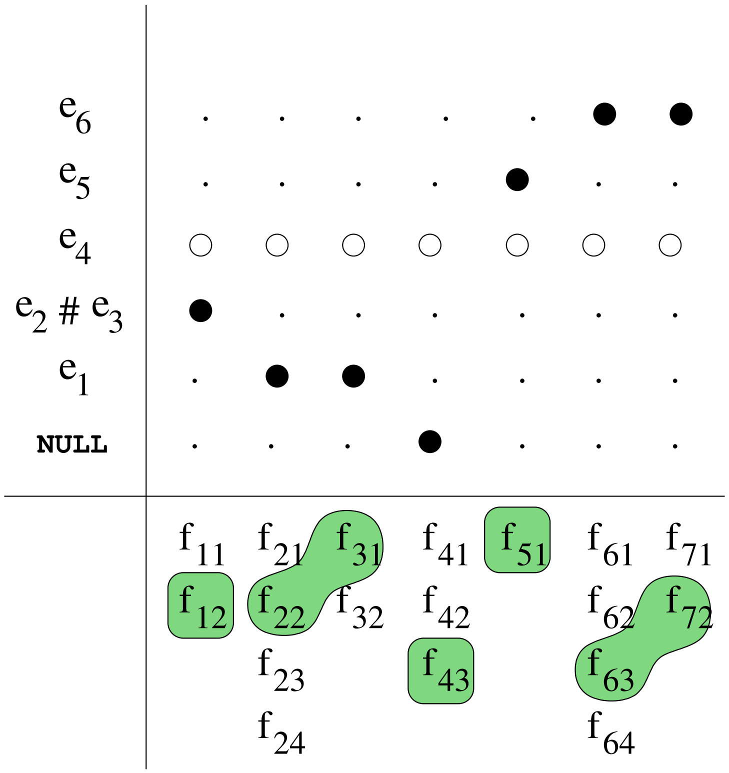
Phrase-based Translation Model

Phrase-based Translation Model

- **Phrase: sequence of consecutive words**

Phrase-based Translation Model

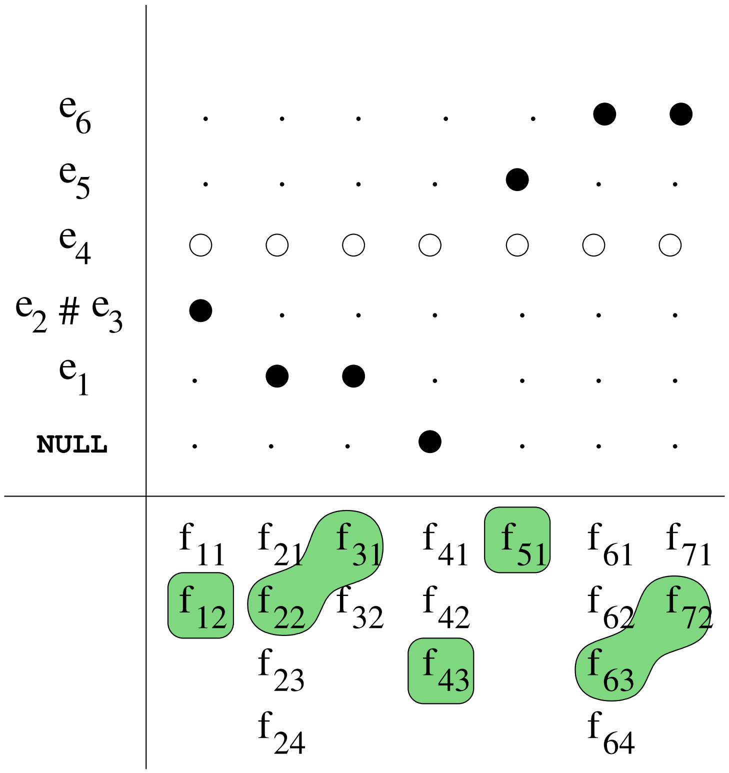
- **Phrase: sequence of consecutive words**
- **Alignment: map between CN and target phrases**
  one word per column aligned with a target phrase



Phrase-based Translation Model

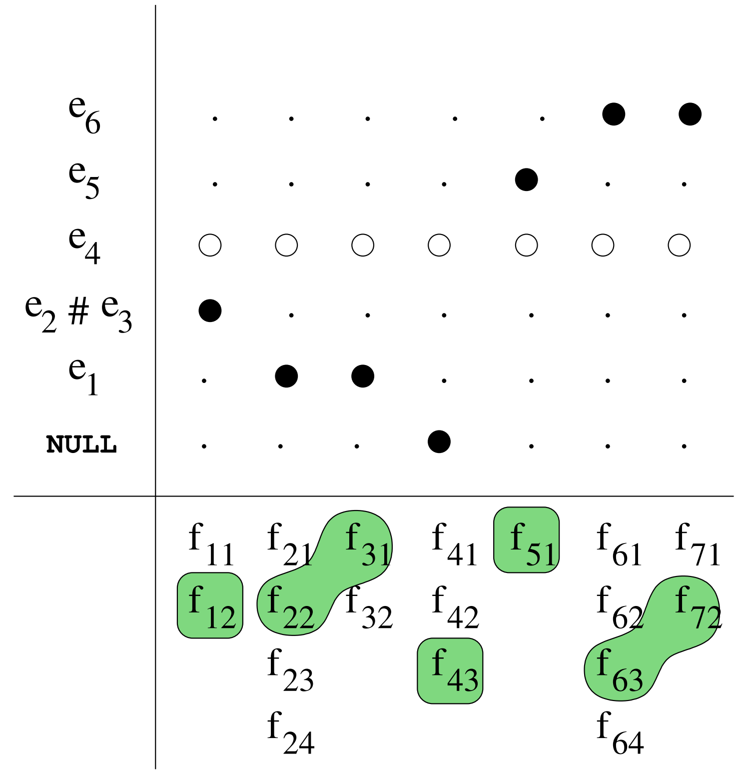
- **Phrase: sequence of consecutive words**
- **Alignment: map between CN and target phrases**
  one word per column aligned with a target phrase
- **Search criterion:**

$$\tilde{\mathbf{e}}^* \approx \arg\max_{\tilde{\mathbf{e}}} \max_{\mathbf{a} \in \mathcal{A}(\mathcal{G}, \tilde{\mathbf{e}})} \Pr(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G})$$



Phrase-based Translation Model

- **Phrase: sequence of consecutive words**
- **Alignment: map between CN and target phrases**
  one word per column aligned with a target phrase



- **Search criterion:**

$$\tilde{\mathbf{e}}^* \approx \arg\max_{\tilde{\mathbf{e}}} \max_{\mathbf{a} \in \mathcal{A}(\mathcal{G},\tilde{\mathbf{e}})} \Pr(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G})$$

- $\Pr(\tilde{\mathrm{e}}, \mathrm{a} \mid \mathcal{G})$ **is a log-linear phrase-based model**

Log-Linear Phrase-based Translation Model

Log-Linear Phrase-based Translation Model

The conditional distribution $\Pr(\tilde{e}, a \mid \mathcal{G})$ is determined through suitable real valued feature functions $h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}), r = 1 \ldots R$, and takes the parametric form:

$$p_{\boldsymbol{\lambda}}(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}) \propto \exp\left\{\sum_{r=1}^{R} \lambda_r h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G})\right\} \tag{1}$$

Log-Linear Phrase-based Translation Model

The conditional distribution $\Pr(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G})$ **is determined through suitable real valued feature functions** $h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}), r = 1 \ldots R$**, and takes the parametric form:**

$$p_{\lambda}(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}) \propto \exp\left\{\sum_{r=1}^{R} \lambda_r h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G})\right\} \quad (1)$$

Feature Functions:

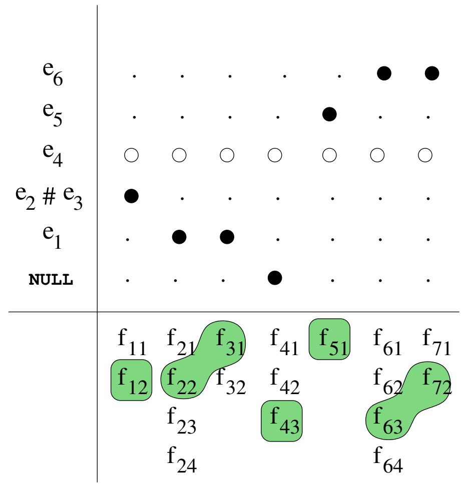
Log-Linear Phrase-based Translation Model

The conditional distribution $\Pr(\tilde{e}, a \mid \mathcal{G})$ is determined through suitable real valued feature functions $h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}), r = 1 \ldots R$, and takes the parametric form:

$$p_\lambda(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}) \propto \exp\left\{\sum_{r=1}^{R} \lambda_r h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G})\right\} \quad (1)$$

Feature Functions:

- **Language model: 3-gram LM**



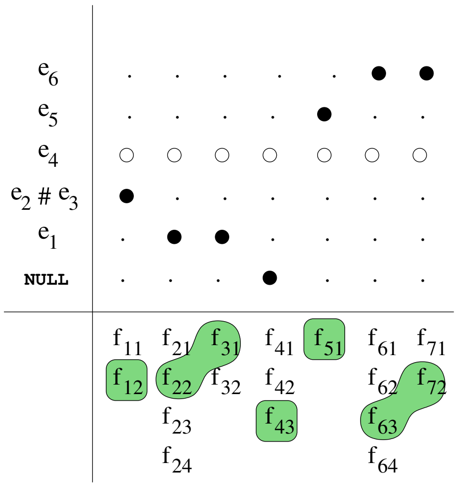
Log-Linear Phrase-based Translation Model

The conditional distribution $\Pr(\tilde{e}, a \mid \mathcal{G})$ **is determined through suitable real valued feature functions** $h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}), r = 1 \ldots R$**, and takes the parametric form:**

$$p_{\lambda}(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}) \propto \exp\left\{\sum_{r=1}^{R} \lambda_r h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G})\right\} \quad (1)$$

Feature Functions:

- **Language model: 3-gram LM**
- **Fertility models: for target phrases and `NULL` word**



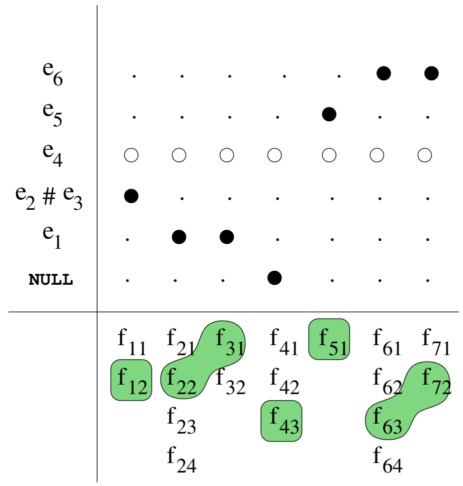
Log-Linear Phrase-based Translation Model

The conditional distribution $\Pr(\tilde{\mathrm{e}}, \mathrm{a} \mid \mathcal{G})$ is determined through suitable real valued feature functions $h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}), r = 1 \ldots R$, and takes the parametric form:

$$p_{\lambda}(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}) \propto \exp\left\{\sum_{r=1}^{R} \lambda_r h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G})\right\} \quad \textbf{(1)}$$

Feature Functions:

- **Language model: 3-gram LM**
- **Fertility models: for target phrases and `NULL` word**
- **Distortion models: reordering of phrases and `NULL` word**



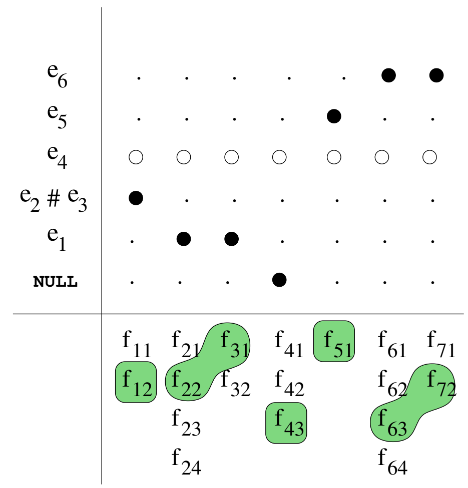
Log-Linear Phrase-based Translation Model

The conditional distribution $\Pr(\tilde{\mathrm{e}}, \mathrm{a} \mid \mathcal{G})$ is determined through suitable real valued feature functions $h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}), r = 1 \ldots R$, and takes the parametric form:

$$p_{\lambda}(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}) \propto \exp\left\{\sum_{r=1}^{R} \lambda_r h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G})\right\} \tag{1}$$

Feature Functions:

- Language model: 3-gram LM
- Fertility models: for target phrases and `NULL` word
- Distortion models: reordering of phrases and `NULL` word
- Lexicon models: phrase-based



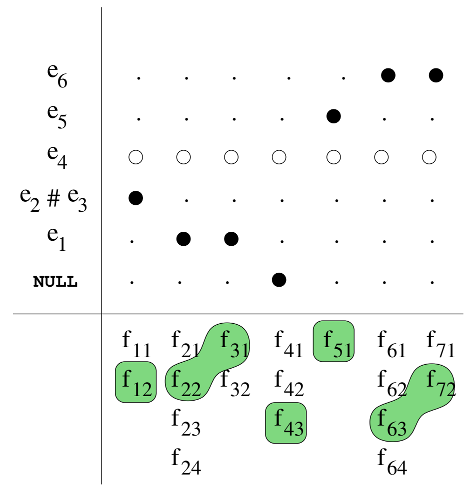
Log-Linear Phrase-based Translation Model

The conditional distribution $\Pr(\tilde{\mathrm{e}}, \mathrm{a} \mid \mathcal{G})$ **is determined through suitable real valued feature functions** $h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}), r = 1 \ldots R$**, and takes the parametric form:**

$$p_{\lambda}(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}) \propto \exp\left\{\sum_{r=1}^{R} \lambda_r h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G})\right\} \quad \textbf{(1)}$$

Feature Functions:

- **Language model: 3-gram LM**
- **Fertility models: for target phrases and `NULL` word**
- **Distortion models: reordering of phrases and `NULL` word**
- **Lexicon models: phrase-based**
- **Likelihood of the path within** $\mathcal{G}$



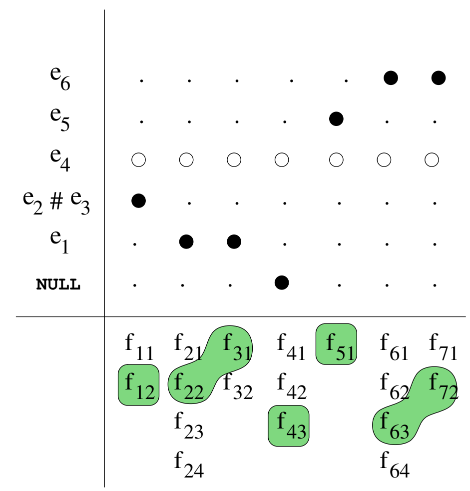
Log-Linear Phrase-based Translation Model

The conditional distribution $\Pr(\tilde{\mathrm{e}}, \mathrm{a} \mid \mathcal{G})$ is determined through suitable real valued feature functions $h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}), r = 1 \ldots R$, and takes the parametric form:

$$p_{\lambda}(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G}) \propto \exp\left\{\sum_{r=1}^{R} \lambda_r h_r(\tilde{\mathbf{e}}, \mathbf{a} \mid \mathcal{G})\right\} \quad \textbf{(1)}$$

Feature Functions:

- **Language model: 3-gram LM**
- **Fertility models: for target phrases and `NULL` word**
- **Distortion models: reordering of phrases and `NULL` word**
- **Lexicon models: phrase-based**
- **Likelihood of the path within $\mathcal{G}$**
- **True length of the path disregarding ϵ-words**



Process for generating a translation hypothesis

Translation

Score:

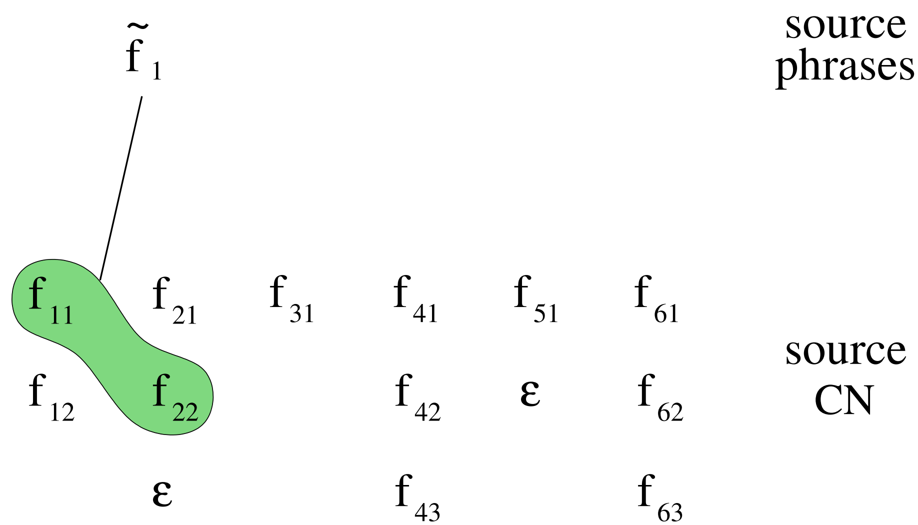
f_{11}	f_{21}	f_{31}	f_{41}	f_{51}	f_{61}	
f_{12}	f_{22}		f_{42}	ε	f_{62}	source CN
	ε		f_{43}		f_{63}	

Process for generating a translation hypothesis

1. choose how many columns
2. choose which consecutive columns
3. choose a word for each column

Translation

Score:

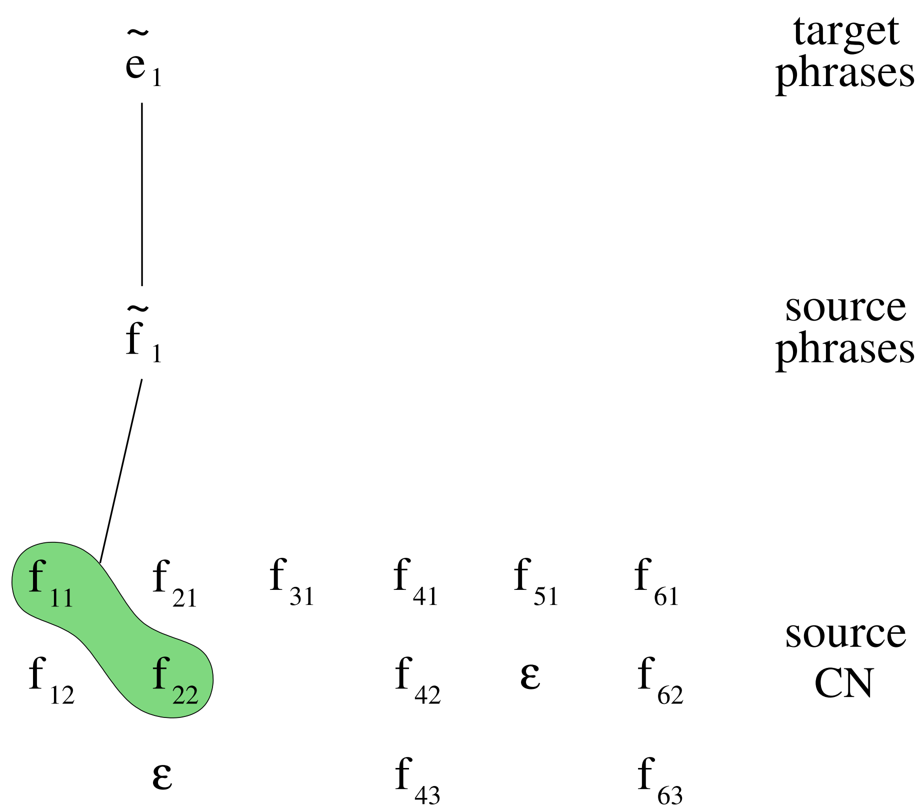


Process for generating a translation hypothesis

1. choose how many columns
2. choose which consecutive columns
3. choose a word for each column
4. choose a target phrase

Translation

Score:

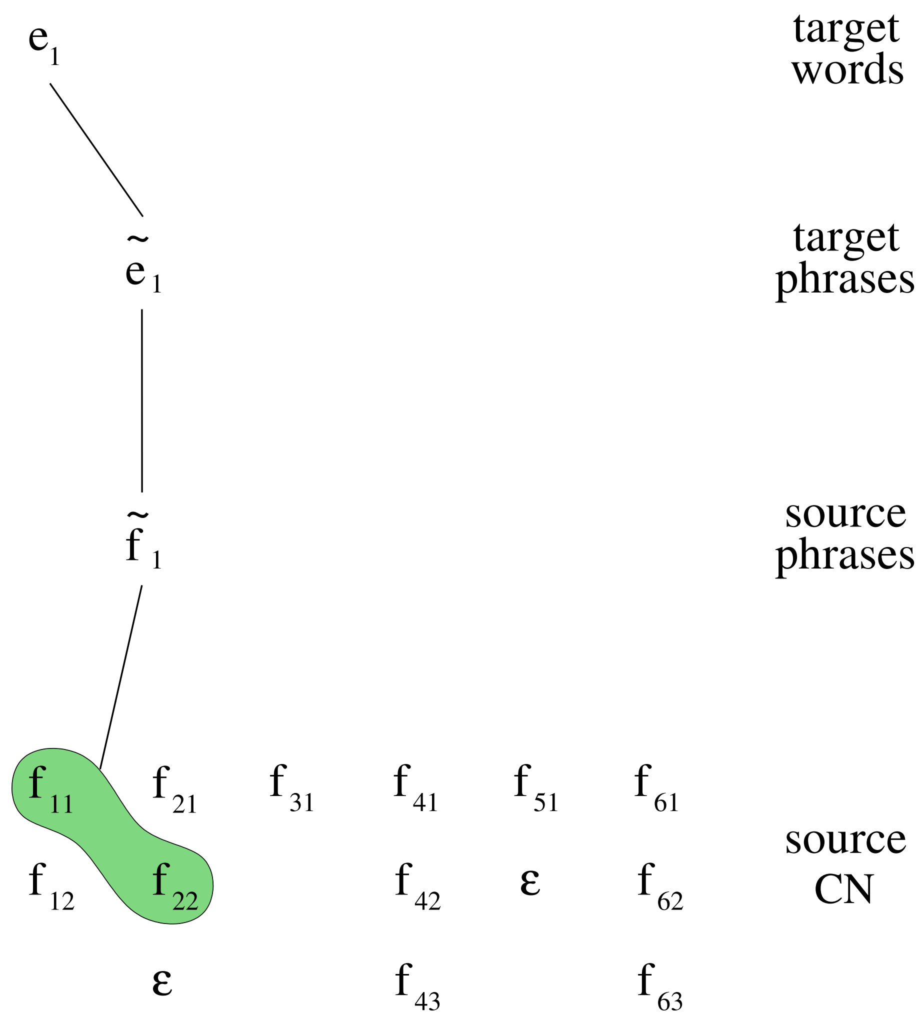


Process for generating a translation hypothesis

1 choose how many columns
2 choose which consecutive columns
3 choose a word for each column
4 choose a target phrase
5 split target phrase into words

Translation e_1

Score:



Process for generating a translation hypothesis

1 choose how many columns

2 choose which consecutive columns

3 choose a word for each column

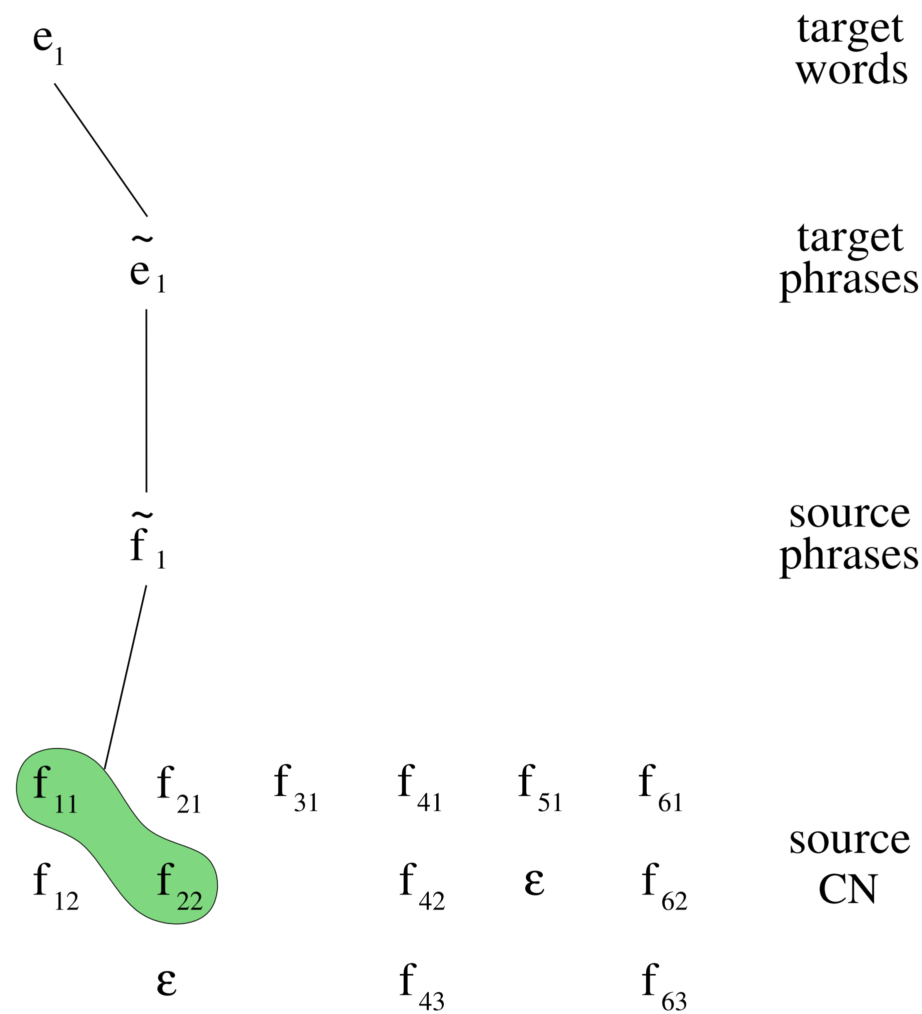
4 choose a target phrase

5 split target phrase into words

6 compute score

Translation e_1

Score: s_1



Process for generating a translation hypothesis

1 choose how many columns

2 choose which consecutive columns

3 choose a word for each column

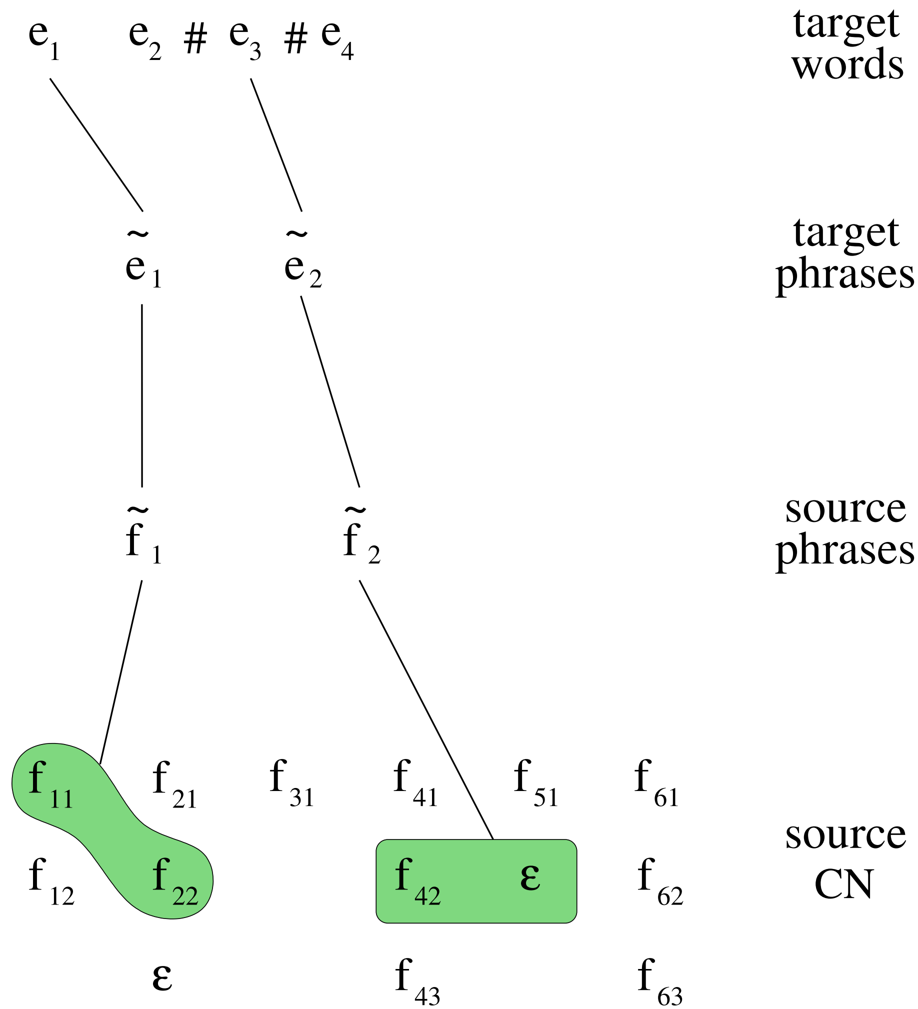
4 choose a target phrase

5 split target phrase into words

6 compute score

Translation e_1 e_2 e_3 e_4

Score: $s_1 + s_2$



Process for generating a translation hypothesis

1 choose how many columns

2 choose which consecutive columns

3 choose a word for each column

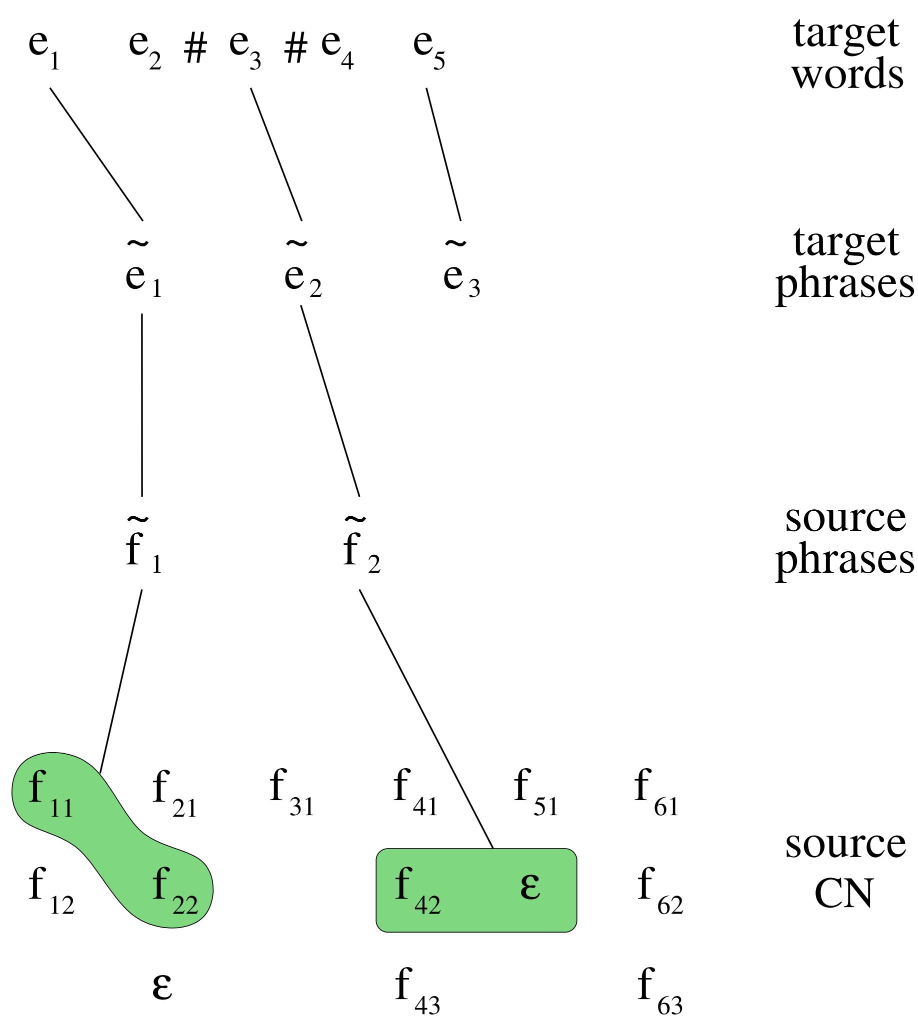
4 choose a target phrase

5 split target phrase into words

6 compute score

Translation e_1 e_2 e_3 e_4 e_5

Score: $s_1 + s_2 + s_3$



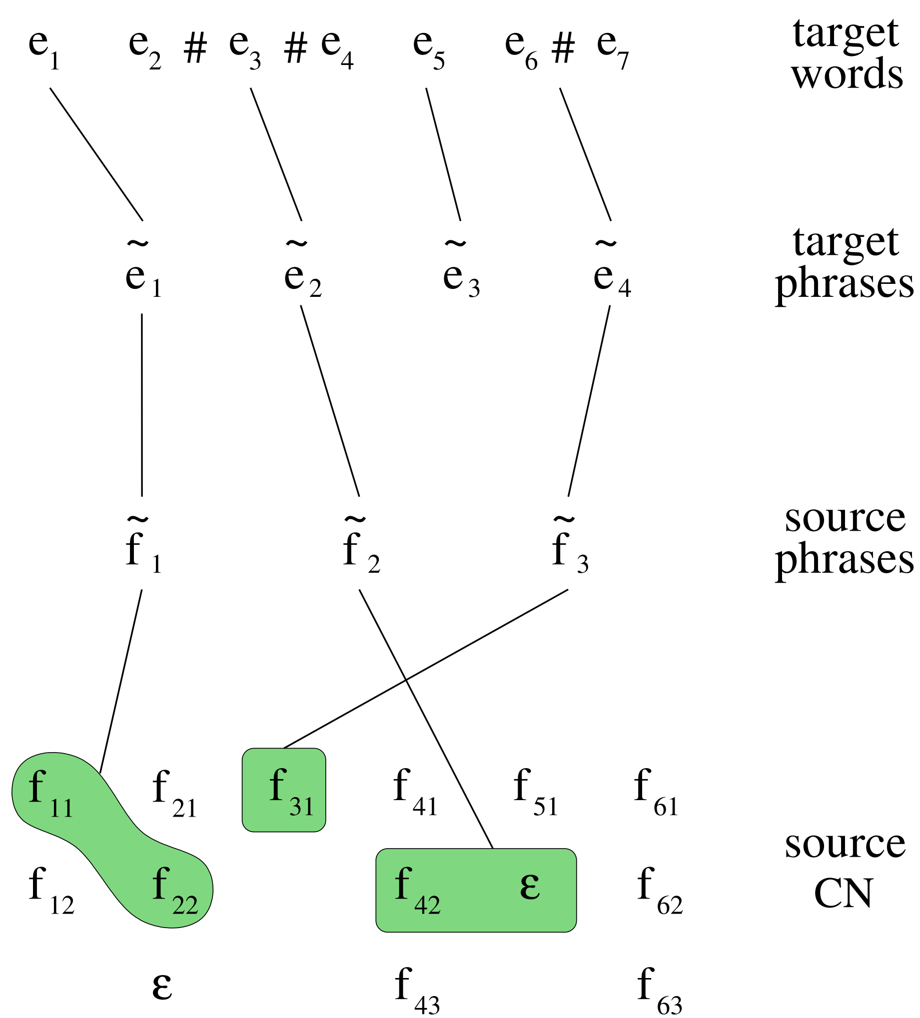
Process for generating a translation hypothesis

1 choose how many columns
2 choose which consecutive columns
3 choose a word for each column
4 choose a target phrase
5 split target phrase into words

6 compute score

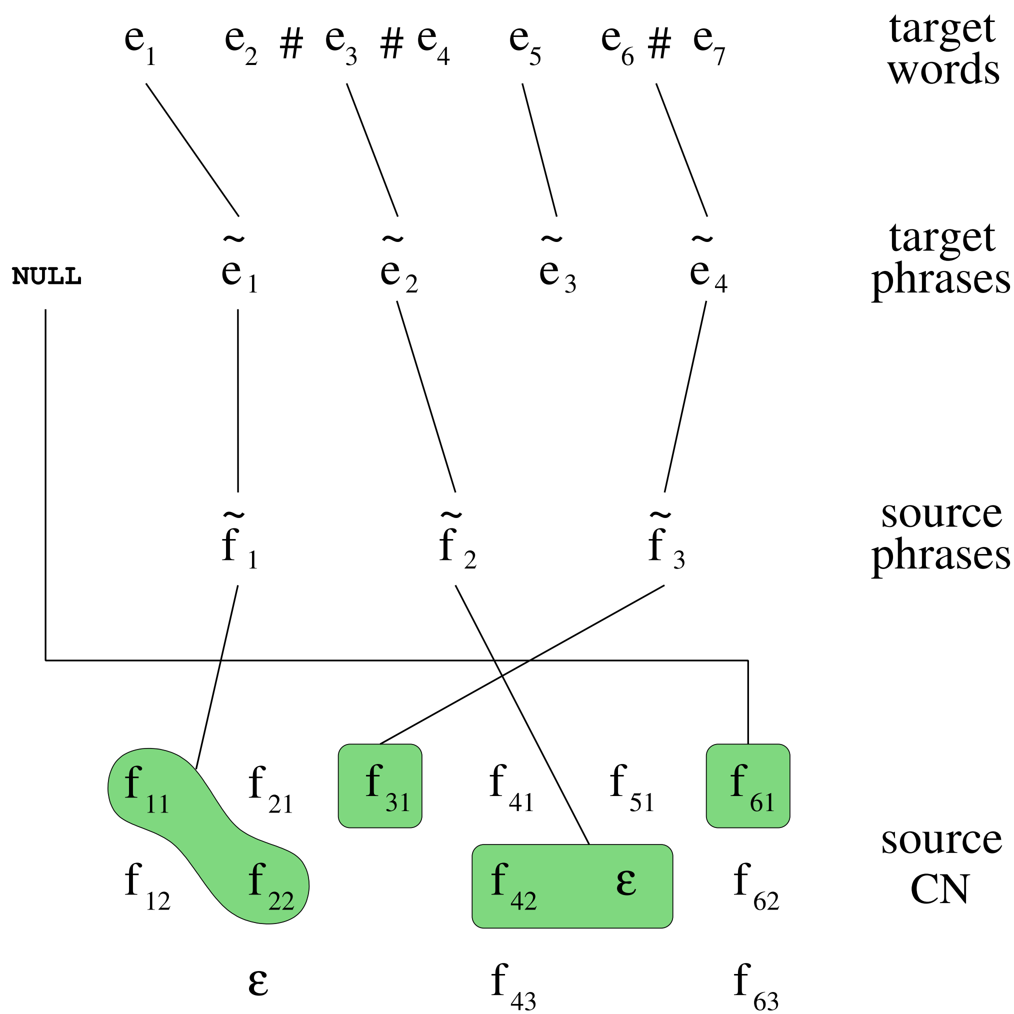
Translation e_1 e_2 e_3 e_4 e_5 e_6 e_7

Score: $s_1 + s_2 + s_3 + s_4$



Process for generating a translation hypothesis

1 choose how many columns
2 choose which consecutive columns
3 choose a word for each column
4 choose a target phrase
5 split target phrase into words
5a not translate words
6 compute score



Score: $s_1 + s_2 + s_3 + s_4 + s_0$

Translation e_1 e_2 e_3 e_4 e_5 e_6 e_7

Decoder

Decoder

- Generative translation process
- Synchronous on output phrases

Decoder

- Generative translation process
- Synchronous on output phrases
- Dynamic programming
- Beam search: deletion of less promising partial translations

Decoder

- Generative translation process
- Synchronous on output phrases
- Dynamic programming
- Beam search: deletion of less promising partial translations
- Reordering constraints: reduction of possible alignments
- Lexicon pruning: no more than 30 translation per phrase
- Confusion network pruning: removal of less confident words

Decoder

- Generative translation process
- Synchronous on output phrases
- Dynamic programming
- Beam search: deletion of less promising partial translations
- Reordering constraints: reduction of possible alignments
- Lexicon pruning: no more than 30 translation per phrase
- Confusion network pruning: removal of less confident words
- Word graph generation: representation of the whole search space
- N-best extraction: multiple translations

N-best-based SLT system

N-best-based SLT system

- relies on a text-based decoder: simplified verison of the CN-based decoder

N-best-based SLT system

- relies on a text-based decoder: simplified verison of the CN-based decoder
- translates separately all N-best transcriptions

N-best-based SLT system

- relies on a text-based decoder: simplified verison of the CN-based decoder
- translates separately all N-best transcriptions
- adds acoustic and source LM scores provided with the N-best transcriptions

N-best-based SLT system

- relies on a text-based decoder: simplified verison of the CN-based decoder
- translates separately all N-best transcriptions
- adds acoustic and source LM scores provided with the N-best transcriptions
- reranks the outputs

Evaluation

- Shared Task T3: integration of ASR and MT
- Input: human, automatic, N-best, Confusion Networks
- Automatic evaluation: BLEU score, case insensitive

	Train		Dev		Test	
Sentences	1,2M		2,643		1,073	
Running words	31M	30M	20K	23K	18.9K	19.3K
Vocabulary	140K	94K	2.9K	2.6K	3.3K	2.8K
best transcription WER	—		11.77%		14.90%	

Results

	DEV				TEST			
	input		BLEU	time	input		BLEU	time
	size	WER			size	WER		
human	1	0	45.78	0.6	1	0	40.84	1.7

Results

	DEV				TEST			
	input		BLEU	time	input		BLEU	time
	size	WER			size	WER		
human	1	0	45.78	0.6	1	0	40.84	1.7
1-bst	1	11.77	40.17	0.6	1	14.60	36.64	2.1

- 10% decrement due to ASR
- comparable to ASR WER

Results

	DEV				TEST			
	input		BLEU	time	input		BLEU	time
	size	WER			size	WER		
human	1	0	45.78	0.6	1	0	40.84	1.7
1-bst	1	11.77	40.17	0.6	1	14.60	36.64	2.1
5-bst	4	8.12	40.63	2.8	5	11.90	36.47	10.5
10-bst	8	6.99	40.83	5.3	9	11.02	36.75	20.4
20-bst	13	6.19	41.03	9.8	16	10.20	36.55	38.9
50-bst	25	5.40	40.85	20.6	34	9.47	36.66	84.2
100-bst	38	5.07	40.87	33.2	56	9.09	36.68	135.3

- 10% decrement due to ASR
- comparable to ASR WER
- few transcriptions
- difficult to improve

Results

	DEV				TEST			
	input		BLEU	time	input		BLEU	time
	size	WER			size	WER		
human	1	0	45.78	0.6	1	0	40.84	1.7
1-bst	1	11.77	40.17	0.6	1	14.60	36.64	2.1
5-bst	4	8.12	40.63	2.8	5	11.90	36.47	10.5
10-bst	8	6.99	40.83	5.3	9	11.02	36.75	20.4
20-bst	13	6.19	41.03	9.8	16	10.20	36.55	38.9
50-bst	25	5.40	40.85	20.6	34	9.47	36.66	84.2
100-bst	38	5.07	40.87	33.2	56	9.09	36.68	135.3
cn-p00	1	11.67	40.30	4.0	1	14.46	36.54	28.4
cn-p50	4	9.42	41.06	5.8	32	11.86	37.14	31.2
cn-p55	13	8.93	41.21	6.3	150	11.32	37.23	34.7
cn-p60	194	8.41	41.24	6.7	1,284	10.71	37.21	37.9
cn-p65	1,359	7.91	41.21	7.4	9,816	10.16	37.05	43.9
cn-p70	15,056	7.53	41.23	27.4	228,461	9.71	37.14	54.6

- 10% decrement due to ASR
- comparable to ASR WER

- few transcriptions
- difficult to improve

- CN slightly better than N-bst
- CN contains more hypotheses
- higher ASR WER
- CN is more efficient

Future plan

Future plan

- generation of richer CNs
  - with lower WER
  - with limited size

Future plan

- generation of richer CNs
  - with lower WER
  - with limited size
- introduction of other features related to input:
  - source LM: reliability of a path

Future plan

- generation of richer CNs
  - with lower WER
  - with limited size
- introduction of other features related to input:
  - source LM: reliability of a path
- experiment on a more difficult task (higher ASR WER)

Future plan

- generation of richer CNs
  - with lower WER
  - with limited size
- introduction of other features related to input:
  - source LM: reliability of a path
- experiment on a more difficult task (higher ASR WER)
- decoding the whole ASR WG

Thanks for your attention!

References

[1] Ney, “Speech Translation: Coupling of Recognition and Translation”. *ICASSP 1999*.

[2] Bangalore and Riccardi, “Stochastic finite-state models for spoken language machine translation”. *Machine Translation*, 17(3), 2002.

[3] Zhang et al., “A Unified Approach in Speech-to-Speech Translation: Integrating Features of Speech Recognition and Machine Translation”. *COLING 2004*.

[4] Casacuberta et al., “Some approaches to statistical and finite-state speech-to-speech translation”. *Computer Speech and Language*, 18, 2004.

[5] Mangu et al., “Finding consensus among words: Lattice-based word error minimization”. *ISCA ECSCT 1999*.

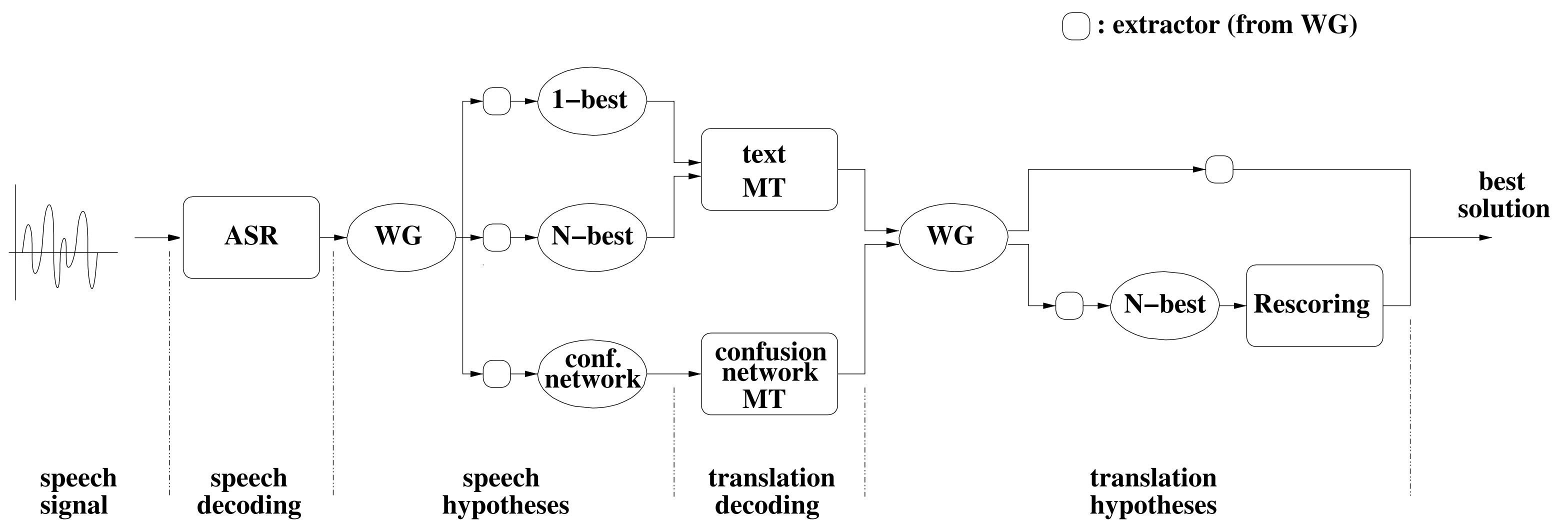
[6] Quan et al., “Integrated n-best re-ranking for spoken language translation”. *Interspeech 2005*.

[7] Cettolo et al., “A look inside the ITC-irst SMT system”. *MT Summit X 2005*.

[8] Federico and Bertoldi, “A word-to-phrase statistical translation model”. Transactions on Speech and Language Processing. 2(2). 2005.

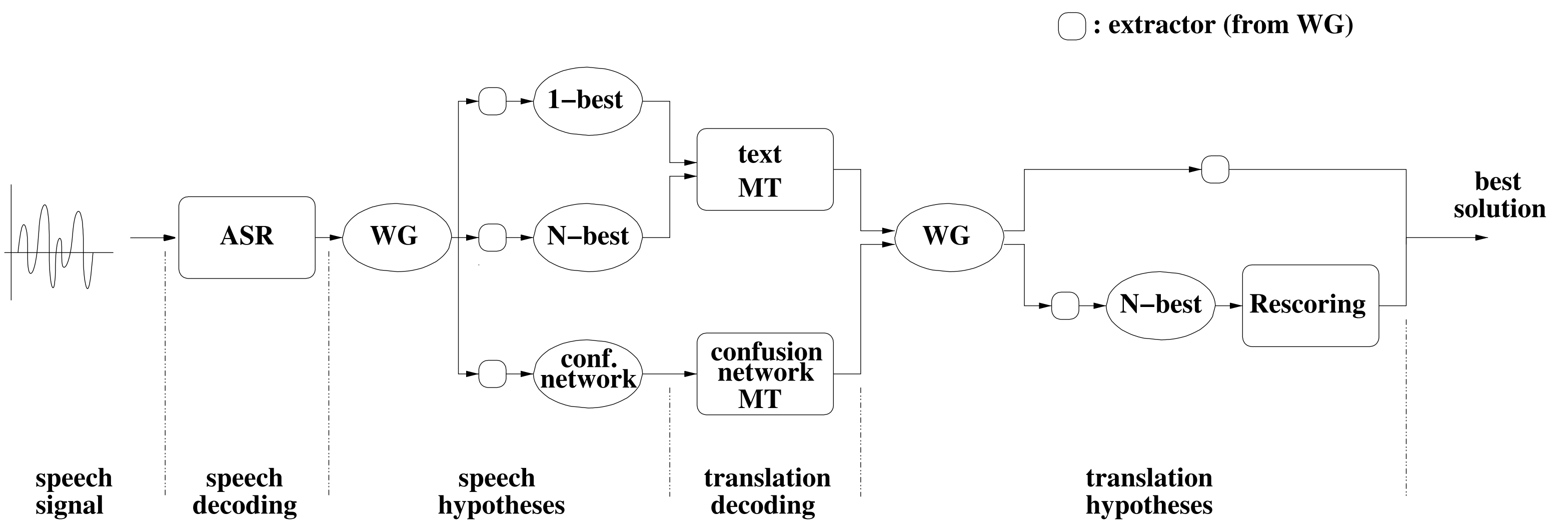
[9] Bertoldi and Federico, “A New Decoder for Spoken Language Translation based on Confusion Networks”. *ASRU 2005*.

ITC-irst SLT system Architecture



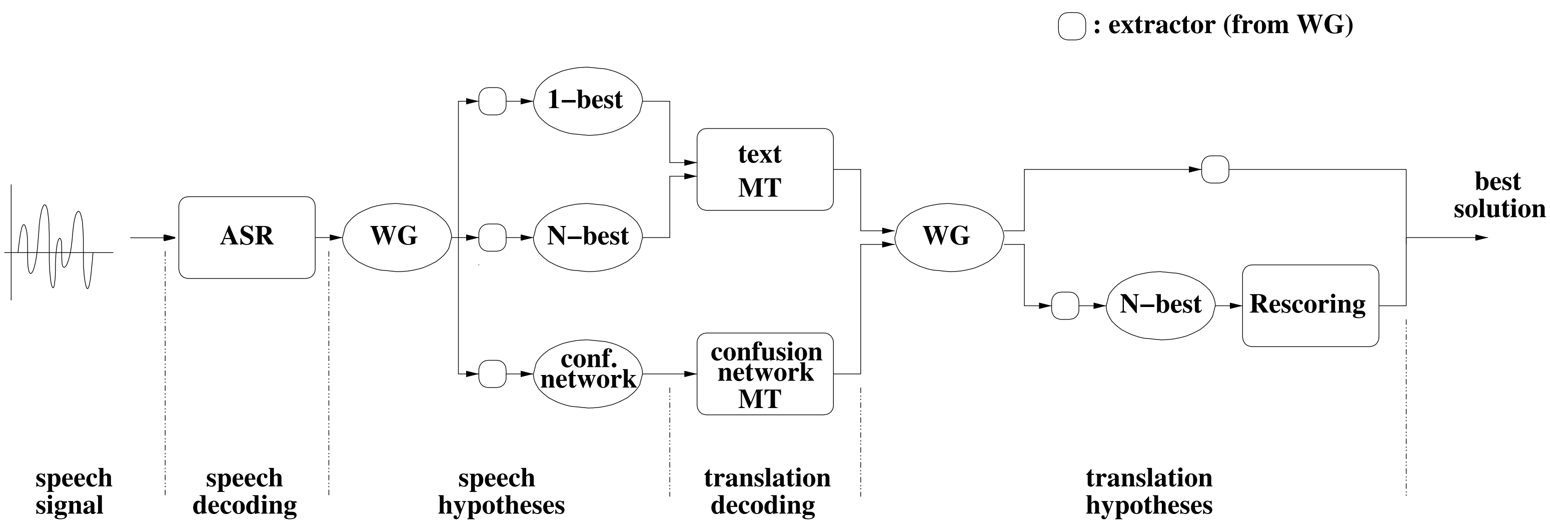
ITC-irst SLT system Architecture

- different input types: text, N-best, Confusion Networks



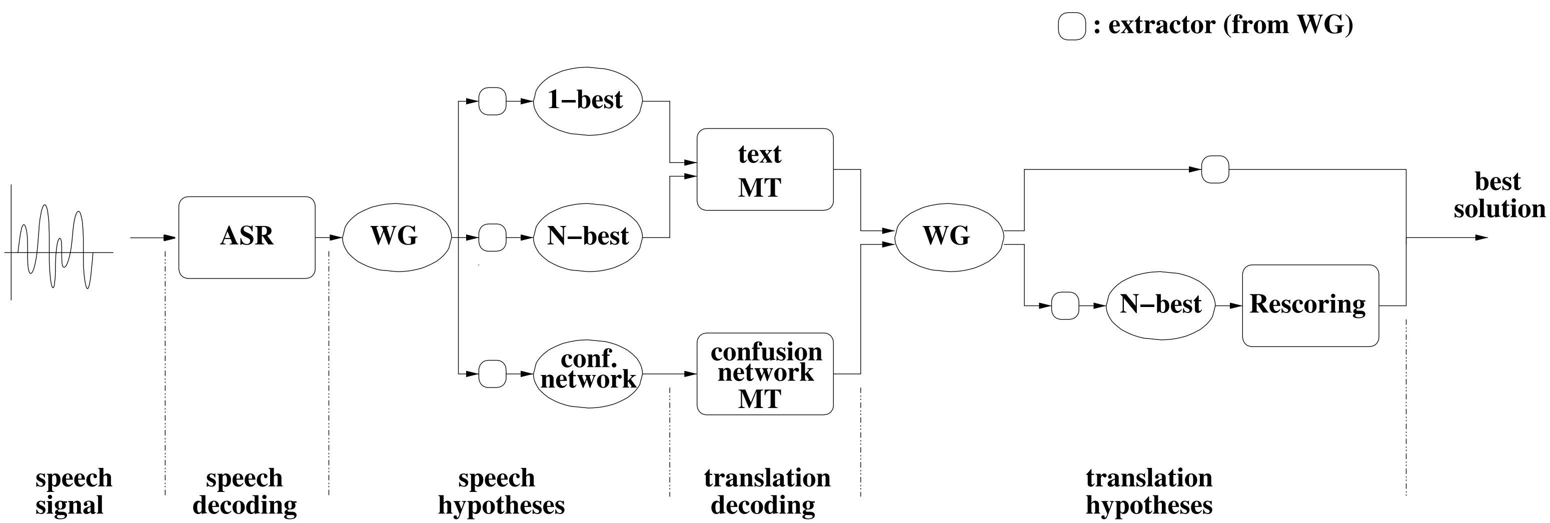
ITC-irst SLT system Architecture

- different input types: text, N-best, Confusion Networks
- two-step decoder



ITC-irst SLT system Architecture

- different input types: text, N-best, Confusion Networks
- two-step decoder
- rescoring with additional features



ITC-irst SLT system Architecture

- different input types: text, N-best, Confusion Networks
- two-step decoder
- rescoring with additional features
- reranking with optimized weights