FONDAZIONE BRUNO KESSLER

Statistical Machine Translation

Marcello Federico
FBK-irst Trento, Italy
Galileo Galilei PhD School –University of Pisa

Pisa, 7-19 May 2008

Part VII: Spoken Language Translation

- **Part 1: Introduction to SLT**
  - A bit of recent history of SLT
  - Statistical Framework of MT
  - Log-linear Phrase-based models
  - Stochastic Finite State Transducers
- **Part 2: Search Architectures for SLT**
  - Integrated vs. Sequential SLT
  - SFST approach
  - N-best List Translation
  - Word-Graph Translation
  - Confusion Network Translation
- **Part 3: TC-STAR Project (FBK's view)**
  - Challenges and Tasks
  - System architecture
  - Progress over time

Credits

- F. Casacuberta, M. Federico, H. Ney, E. Vidal, "Recent Efforts in Spoken Language Translation", *IEEE Signal Processing Magazine*, May, 2008.
- N. Bertoldi, R. Zens, M. Federico, W. Shen, "Efficient Speech Translation through Confusion Network Decoding", Accepted in *IEEE Transactions on Audio, Speech, and Language Processing*, 2008.

PART 1: Introduction To SLT

Spoken Language Translation

Translation from speech input is **more difficult** than translation from text input:

- many *styles* and *genres:*
  formal read speech, unplanned speeches, interviews,
  spontaneous conversations, ...
- *less controlled* language:
  relaxed syntax, spontaneous speech phenomena
- automatic speech recognition is prone to *errors*:
  possible corruption of syntax and meaning

SLT rises two relevant issues:

- training data: most language resources are for written language
- **integration of ASR and MT**: propagation of errors

A Bit of Recent History on SLT

- 1997: **ACL Workshop on SLT**
  – *communication oriented* rather than translation oriented (e.g. interlingua)
  – interactive SLT exploiting dialogue structure
  – limited domains: traveling, appointment scheduling
  – data collected in the laboratory
  – statistically and linguistically motivated approaches
- 1999:**C-STAR II Multilingual Demonstator**
  – *interlingua-based*, data-driven translation, traveling domain
- 2000: **Verbmobil Demonstrator**
  – *statistical approach* vs. rule-based SLT, traveling domain
- 1997-2004: **EU Projects: EuTrans, Nespole!, PF-STAR**
  – *integration of ASR and MT:* SFST, log-linear models, traveling domain
- 2004-:**GALE and TC-STAR projects**
  – large vocabulary ASR technology (from laboratory speech to *found speech*)
  – *unconstrained SLT* (news and political speeches)

Interlingua-Based Translation (C-STAR, 1999)

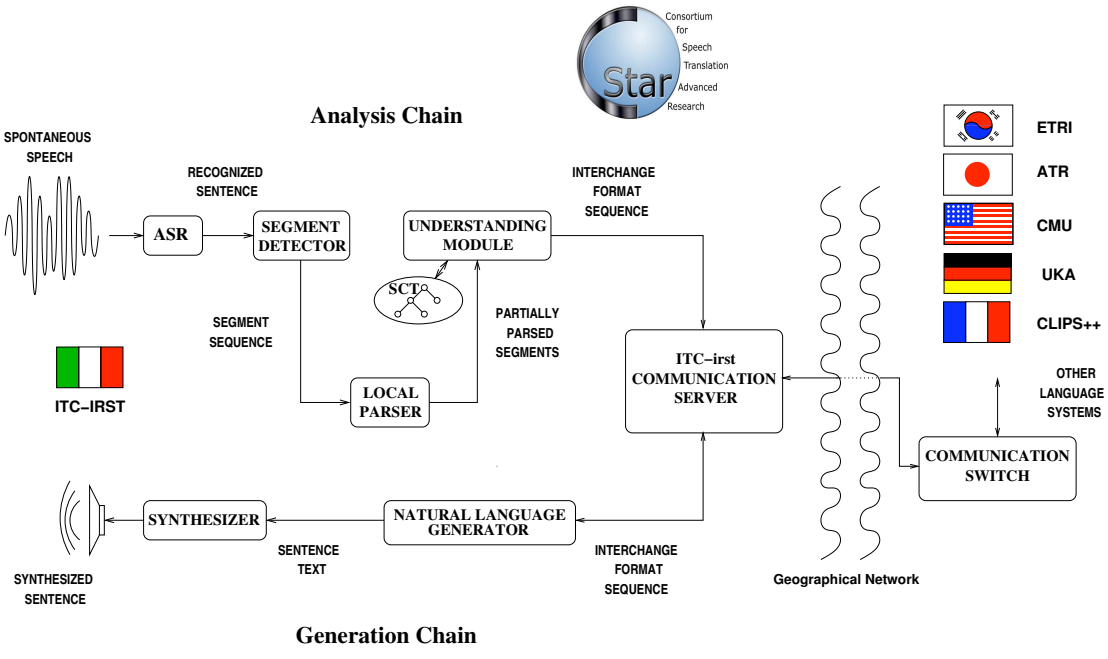
- S[1] : I'm arriving on june sixth
- I: `give-information+temporal+arrival (who=I, time=(june, md6))`
- T: my arrival time is sixth of june

- S: no that's not necessary
- I: `negate`
- T: no

- S: and i was wondering what you have in the way of rooms available during that time
- I: `request-information+availability+room (room-type=question)`
- T: what kind of rooms are available?

[1] S: speech (English), I: Interlingua, T: translation (English)

Interlingua-Based Translation (C-STAR, 1999)



Spoken Language Translation

SLT combines the difficulties of two problems:

- ASR: conversion from speech to text
- MT: translation from source language to target language

$$\begin{array}{lll} \text{speech signal} \rightarrow & \text{source text} & \rightarrow \text{target text} \\ \mathbf{x} \rightarrow & \mathbf{f} & \rightarrow \mathbf{e} \end{array}$$

Speech misses important semantic/syntactic cues of written language:

- paragraph and sentence delimiters
- punctuation marks and capitalized words

Input is just a stream of uttered words.

Statistical Framework of SLT

Statistical decision theory suggest to use the *Bayes decision rule* and select the $\hat{\mathbf{e}}(\mathbf{x})$ with the highest *posterior probability*:

$$\begin{aligned} \mathbf{x} \rightarrow \hat{\mathbf{e}}(\mathbf{x}) &= \underset{\mathbf{e}}{\operatorname{argmax}} \{p(\mathbf{e}|\mathbf{x})\} && (1) \\ &= \underset{\mathbf{e}}{\operatorname{argmax}} \{\sum_{\mathbf{f}} p(\mathbf{e},\mathbf{f}|\mathbf{x})\} && (2) \\ &= \underset{\mathbf{e}}{\operatorname{argmax}} \{\sum_{\mathbf{f}} p(\mathbf{e}|\mathbf{f}) \cdot p(\mathbf{f}|\mathbf{x})\} && (3) \\ &\cong \underset{\mathbf{e}}{\operatorname{argmax}} \{\max_{\mathbf{f}} \{p(\mathbf{e}|\mathbf{f}) \cdot p(\mathbf{f}|\mathbf{x})\}\} && (4) \end{aligned}$$

We applied the following *approximations*:

- (3) given $\mathbf{f}$, $\mathbf{x}$ does not provide additional info about $\mathbf{f}$
- (4) maximum approximation, which is widely used in ASR.

Statistical Framework of SLT

With simple manipulations we can derive the following equivalent criteria:

- **Chain SLT model**

$$\mathbf{x} \rightarrow (\hat{\mathbf{e}}, \hat{\mathbf{f}})(\mathbf{x}) \quad = \quad \underset{\mathbf{e},\mathbf{f}}{\operatorname{argmax}} \left\{ p(\mathbf{e}|\mathbf{f}) \cdot p(\mathbf{f}|\mathbf{x}) \right\} \tag{5}$$

- **Joint SLT model**

$$\mathbf{x} \rightarrow (\hat{\mathbf{e}}, \hat{\mathbf{f}})(\mathbf{x}) \quad = \quad \underset{\mathbf{e},\mathbf{f}}{\operatorname{argmax}} \left\{ p(\mathbf{e},\mathbf{f}) \cdot p(\mathbf{x}|\mathbf{f}) \right\} \tag{6}$$

SLT is the problem of combining and integrating ASR and MT:

- to achieve *optimal performance*
- under reasonable *computational requirements*

Statistical Framework of SLT

Optimization through the **Bayes decision rule is computationally challenging**

- Unlike ASR, MT requires *non-monotonic* alignments to allow for different word orders in source and target sentence.
- In the general case, the search algorithm has to consider a large number of different *word re-orderings*

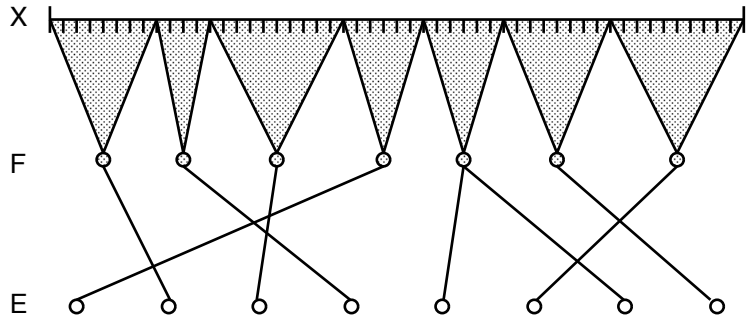


Figure 1: Illustration of the search Problem in SLT.

Sequential Approximation

In a first approximation, we can try to decompose the SLT task into ASR and MT with the associated decision rules:

$$\text{ASR:} \quad \mathbf{x} \rightarrow \hat{\mathbf{f}}(\mathbf{x}) = \underset{\mathbf{f}}{\operatorname{argmax}} \{p(\mathbf{f}|\mathbf{x})\} \tag{7}$$

$$\text{MT:} \quad \hat{\mathbf{f}}(\mathbf{x}) \rightarrow \hat{\mathbf{e}}(\hat{\mathbf{f}}(\mathbf{x})) = \underset{\mathbf{e}}{\operatorname{argmax}} \{p(\mathbf{e}|\hat{\mathbf{f}}(\mathbf{x}))\} \tag{8}$$

which can be interpreted as a *sequential approximation* to the SLT optimization problem:

$$\mathbf{x} \rightarrow \hat{\mathbf{e}}(\mathbf{x}) = \underset{\mathbf{e}}{\operatorname{argmax}} \left\{ \underset{\mathbf{f}}{max} \{p(\mathbf{e}|\mathbf{f}) \cdot p(\mathbf{f}|\mathbf{x})\} \right\} \tag{9}$$

$$\cong \arg \underset{\mathbf{e}}{max} \left\{ p\big(\mathbf{e}| \underset{\mathbf{f}}{\operatorname{argmax}}\, p(\mathbf{f}|\mathbf{x})\big) \right\} \tag{10}$$

Log-Linear Phrase Modeling

To model the posterior $p(\mathbf{e}|\mathbf{f})$, we first introduce an *explicit normalization*:

$$p(\mathbf{e}|\mathbf{f}) = \frac{Q(\mathbf{e},\mathbf{f})}{\sum_{\tilde{\mathbf{e}}} Q(\tilde{\mathbf{e}},\mathbf{f})} \tag{11}$$

$$Q(\mathbf{e},\mathbf{f}) := \underset{\mathbf{b}}{max}\ Q(\mathbf{e},\mathbf{f};\mathbf{b}) \tag{12}$$

- The *hidden variable* $\mathbf{b}$ summarizes phrase segmentation and re-ordering.
- $Q(\mathbf{e},\mathbf{f};\mathbf{b})$ embeds probabilistic dependencies between $\mathbf{e},\mathbf{f}$ and $\mathbf{b}$ under form of *feature functions* denoted by $h_m(\mathbf{e},\mathbf{f};\mathbf{b}), m = 1,...,M$
- We assume a *log-linear relationship* between $Q(\mathbf{e},\mathbf{f};\mathbf{b})$ and $h_m(\mathbf{e},\mathbf{f};\mathbf{b})$:

$$\log\ Q(\mathbf{e},\mathbf{f};\mathbf{b}) = \sum_m \lambda_m h_m(\mathbf{e},\mathbf{f};\mathbf{b}) \tag{13}$$

with *log-linear parameters* $\lambda_m, m = 1,...,M$.

Log-Linear Phrase Modeling

To find the target sentence $\hat{\mathbf{e}}(\mathbf{f})$ for a given source sentence $\mathbf{f}$, the **Bayes decision rule can be re-written** in this log-linear framework as follows:

$$\mathbf{f} \rightarrow \hat{\mathbf{e}}(\mathbf{f}) = \underset{\mathbf{e}}{\operatorname{argmax}} \left\{ p(\mathbf{e}|\mathbf{f}) \right\} \quad (14)$$

$$= \underset{\mathbf{e}}{\operatorname{argmax}} \left\{ \max_{\mathbf{b}} Q(\mathbf{e}, \mathbf{f}; \mathbf{b}) \right\} \quad (15)$$

$$= \underset{\mathbf{e}}{\operatorname{argmax}} \left\{ \max_{\mathbf{b}} \left\{ \sum_{m} \lambda_m h_m(\mathbf{e}, \mathbf{f}; \mathbf{b}) \right\} \right\} \quad (16)$$

- The attractive property of this log-linear approach is that we **do not** have to **worry about the normalization** of the various probabilistic dependencies.
- Typically, the **log-linear parameters are optimized for optimum translation accuracy** on a development set [Och and Ney, 2002].

Log-Linear Phrase Modeling: Features

Most relevant features:

- The relative frequency of the phrase pair $(\tilde{e}, \tilde{f})$.
- Lexicon probabilities of the IBM models or the relative frequencies of the (target,source) pairs (e, f) within a phrase-pair
- Target language model (word or class-based n-grams)
- Phrase re-ordering features

Log-Linear Phrase Modeling: Search

Partial translation hypotheses are build *bottom-to-top* over the source positions:

- **Alignments must cover all positions of f exactly once**
  (cf. TSP constraint: each city has to be visited exactly once)

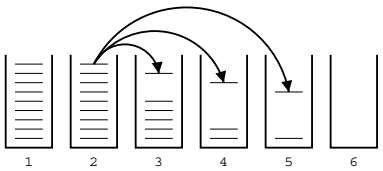
- **Alternatives are explored and scored along different directions**:
  – possible *segmentations* of input into phrases
  – possible *translations* of source phrases
  – possible *re-ordering* of source phrases
- **Search algorithms**: A^* and *DP beam search* [Och and Ney, 2003, Tillmann and Ney, 2003, Koehn et al., 2003, Federico and Bertoldi, 2005]

- **Approximations**: limited word-reordering and translation options

Log-Linear Phrase Modeling: Moses Decoder

- **Phrase-based decoding steps:**
  – **cover** some not yet covered consecutive words (*span*)
  – **retrieve** phrase-translations for the span
  – **compute** translation, distortion and target language models

- **Multi-stack decoder**:
  – theories stored according to the coverage size
  – synchronous on the coverage size
- **DP recombination** of similar partial translations
- **Beam search**:
  – deletion of less promising partial translations:
  – histogram and threshold pruning
- **Distortion limit**: reduction of possible alignments
- **Lexicon pruning**: limit the amount of translation options per span



Stochastic Finite State Transducers

Another way to deal with statistical MT is to compute the optimization (14) via the *joint probability* $p(\mathbf{f}, \mathbf{e})$:

$$\mathbf{f} \rightarrow \hat{\mathbf{e}}(\mathbf{f}) = \underset{\mathbf{e}}{\operatorname{argmax}} \left\{ \frac{p(\mathbf{e}, \mathbf{f})}{p(\mathbf{f})} \right\} \quad (17)$$

$$= \underset{\mathbf{e}}{\operatorname{argmax}} \left\{ p(\mathbf{e}, \mathbf{f}) \right\} \quad (18)$$

- *Stochastic finite-state transducers* (SFSTs) can be used to model $p(\mathbf{e}, \mathbf{f})$ [Casacuberta and Vidal, 2004].
- A SFST is composed of a set of *states* (with an initial state) and a set of *transitions* between pairs of states.
- **Transitions are labeled with source-target phrase-pairs** $(\tilde{f}, \tilde{e})$, and are weighted with a probability.
- Finally, each state has assigned a probability to be a final state.

Stochastic FST

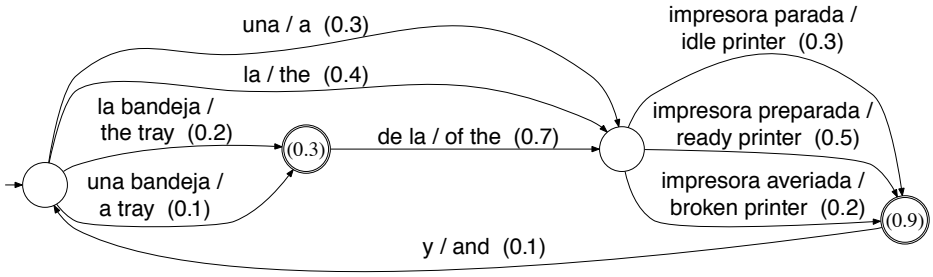


Figure 2: An example of SFST for Spanish to English translation.

$$p(\mathbf{e}, \mathbf{f}) = \sum_{\pi} p(\mathbf{e}, \mathbf{f}, \pi) \quad (19)$$

- $p(\mathbf{e}, \mathbf{f}, \pi)$ is the probability that $(\mathbf{e}, \mathbf{f})$ is generated through a *path* π
- π is a sequence of transitions in the SFST that matches $\mathbf{e}$ and $\mathbf{f}$.
- $p(\mathbf{e}, \mathbf{f}, \pi)$ is computed as the product of all transition probabilities in π

Stochastic FST

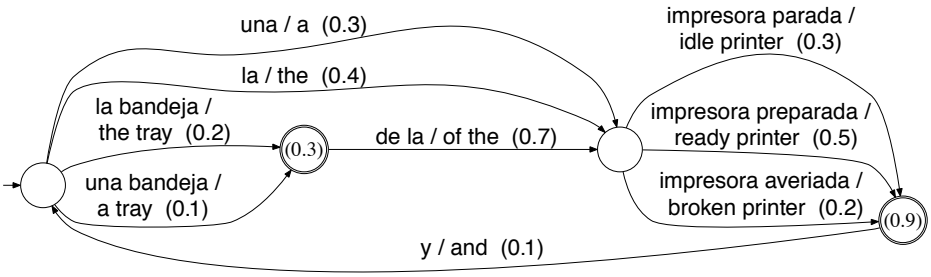


Figure 3: An example of SFST for Spanish to English translation.

$$p(\mathbf{e}, \mathbf{f}) \quad = \quad \sum_{\pi} p(\mathbf{e}, \mathbf{f}, \pi) \tag{20}$$

- **Hidden variable π in SFST is similar to b in log-linear model.**
- Paths π embed exactly phrase segmentation and alignment information.
- Therefore, SFSTs are implicit *phrase based* models.

Stochastic FST: Properties

- **SFSTs are intrinsic monotonic models**, but arbitrary finite-length word reordering can be properly accommodated into the source/target phrases associated to the transitions.
- **However, the size of the SFST can grow very fast if many long-span reordering patterns have to be modeled!**
- A SFST has stochastic source and target regular languages embedded that are obtained, respectively, by dropping the target or the source symbols from all transitions
- SFSTs are special cases of *weighted finite-state transducers*, which have been used to explicitly implement statistical phrase-based [Kumar et al., 2006, Mathias and Byrne, 2006] and word-based models [Knight and Al-Onaizan, 1998].

Stochastic FST: Training

There are several techniques to learn a SFST a parallel corpus.

With the GIATI [2] technique [Casacuberta and Vidal, 2007]:

1. compute word alignments on the parallel corpus
2. compute bilingual segments in a similar way as for phrase-based models
3. estimate a smoothed n-gram LM for these strings and represent it by as a finite-state network;
4. convert the "bilingual symbols" on the transitions back into source/target words.

A similar approach was independently proposed in [Bangalore and Riccardi, 2003]. In [Mariño et al., 2006] this kind of bilingual n-gram modeling is placed within the log-linear framework.

[2] Grammatical Inference and Alignments for Transducer Inference

Stochastic FST: Search

For computational reasons **we apply the maximum approximation**.

- *Viterbi score of a translation*: maximize over all the paths that generate $(\mathbf{e}, \mathbf{f})$

$$p(\mathbf{e}, \mathbf{f}) \approx \max_{\pi} p(\mathbf{e}, \mathbf{f}, \pi) \; . \tag{21}$$

- this score can be computed very efficiently by *DP beam search*
- $\hat{\mathbf{e}}(\mathbf{f})$ for $\mathbf{f}$ is determined easily once the optimal π is computed.

This efficient and effective search strategy makes **SFST attractive at least for not very large (and speech-input) applications**.

Part 2: Search Architectures for SLT

Search Architectures for SLT

There are two basic approaches for the joint modelling of ASR and MT:

- *Integrated architecture*, based on a SFST approximation of

$$\mathbf{x} \rightarrow (\hat{\mathbf{e}}, \hat{\mathbf{f}})(\mathbf{x}) \quad = \quad \underset{\mathbf{e},\mathbf{f}}{\operatorname{argmax}} \left\{ p(\mathbf{e}, \mathbf{f}) \cdot p(\mathbf{x}|\mathbf{f}) \right\} \tag{22}$$

- *Decoupled architecture*, based on a log-linear approximation of

$$\mathbf{x} \rightarrow (\hat{\mathbf{e}}, \hat{\mathbf{f}})(\mathbf{x}) \quad = \quad \underset{\mathbf{e},\mathbf{f}}{\operatorname{argmax}} \left\{ p(\mathbf{e}|\mathbf{f}) \cdot p(\mathbf{f}|\mathbf{x}) \right\} \tag{23}$$

SFST Integrated Models

- **HMMs are used to model $p(\mathbf{x}|\mathbf{f})$ and a SFST to model $p(\mathbf{e},\mathbf{f})$.**
- Computation of $p(\mathbf{x}|\mathbf{f})$ considers *segmentations of* $\mathbf{x}$ to match words of $\mathbf{f}$:

$$p(\mathbf{x},\sigma|\mathbf{f}) \quad \propto \quad p(\tilde{x}_1^J|f_1^J) = \prod_j p(\tilde{x}_j|f_j) \ , \tag{24}$$

  – $\sigma = \tilde{x}_1,\ldots,\tilde{x}_J$ is the segmentation of $\mathbf{x}$ into J words
  – $p(\tilde{x}_j|f_j)$ is the acoustic likelihood of f_j computed by the HMM
- **Maximization over (e,f) is approximated by searching over all paths in the SFST and segmentations in x**:

$$p(\mathbf{f},\mathbf{e})\cdot p(\mathbf{x}|\mathbf{f}) \quad \propto \quad \max_{\pi,\sigma}\{p(\mathbf{f},\mathbf{e},\pi)\cdot p(\mathbf{x},\sigma|\mathbf{f})\} \ . \tag{25}$$

- Such a computation can be carried out by the *Viterbi search algorithm*.

Log-Linear Decoupled Models

Sequential approximation suffers from the *accumulation of ASR and MT errors*.

- **Idea: optimize jointly over a significant subset of ASR hypotheses**
- **Assumption**: $\mathcal{H}(\mathbf{x})$ should likely contain more accurate transcriptions

$$\mathbf{x} \ \rightarrow \ \hat{\mathbf{e}}(\mathbf{x}) \quad = \quad \operatorname*{argmax}_{\mathbf{e}} \left\{ \max_{\mathbf{f}\in\mathcal{H}(\mathbf{x})} Q(\mathbf{e},\mathbf{f};\mathbf{x}) \right\} \tag{26}$$

- The (not normalized) *joint probability* $Q(\mathbf{e},\mathbf{f};\mathbf{x})$ is represented by:

$$Q(\mathbf{e},\mathbf{f};\mathbf{x}) \quad = \quad Q(\mathbf{e},\mathbf{f})^{\alpha_1}\cdot p(\mathbf{f})^{\alpha_2}\cdot p(\mathbf{x}|\mathbf{f})^{\alpha_3} \tag{27}$$

  – $Q(\mathbf{e},\mathbf{f})$ defined in Eq.12 and exponents
  – α_i are optimized on a development set for optimum performance.
- **Difficulty of the search problem** depends on the *size and structure* of $\mathcal{H}(\mathbf{x})$

N-best List Translation

Let $\mathcal{H}(\mathbf{x})$ be list of *N-best ASR outputs*: $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N$

- **Pro: easy to integrate into the log-linear framework** [Zhang et al., 2004]:
  1- Compute M-best translations for each N-best ASR output $\mathbf{f}_n$
  2- Re-score all $N \times M$ translations including ASR scores
  3- Output the top ranking translation

- **Cons: computation increases by factor N**:
  – does not exploit *redundancies* in N-best lists
  – accuracy of N-best lists increases slowly with N

- **Optimal value of N depends on several variables**:
  – difficulty of the ASR task (tricky relationship)
  – length of the input (the longer the more alternatives are needed)
  – setting of ASR beam search (tricky relationship)

Example of ASR N-best List

Source: European Parliament Plenary Sessions

1	as his farewell appearances like iran into his seventies
2	as his farewell appearances iran into his seventies
3	as his farewell appearances on into his seventies
4	as his farewell appearances on into the seventies
5	as his farewell appearances drawn into his seventies
6	as his farewell appearances like iran into the seventies
7	as his farewell appearances iran into the seventies
8	as his farewell appearances drawn into the seventies
9	as his farewell appearances and on into his seventies
10	has his farewell appearances like iran into his seventies
11	as his farewell appearances like enron into his seventies
12	as his farewell appearances like iran in the seventies
13	has his farewell appearances iran into his seventies
15	as his farewell appearances iran in the seventies
16	as his farewell appearances and on into the seventies
17	as his farewell appearances tag and on into his seventies
18	as his farewell appearances gone into his seventies
19	as his farewell appearances on in the seventies
20	as his farewell appearances in iran into his seventies
21	as his farewell appearances like and on into his seventies
22	as his farewell appearances stagger on in the seventies
23	as his farewell appearances stagger on into his seventies
24	as his farewell appearances here on into his seventies
25	has his farewell appearances on into his seventies
26	as his farewell appearances tag and on into the seventies
27	as his farewell appearances like iran and the seventies
29	as his farewell appearances here on in the seventies
30	as his farewell appearance as iran into his seventies
31	as his farewell appearances like and on into the seventies
32	as his farewell appearances that iran into his seventies
33	as his farewell appearances i go on into his seventies
34	as his farewell appearances stagger on into the seventies
35	as his farewell appearances in on into his seventies
36	as his farewell appearances by iran into his seventies
37	as his farewell appearances here on into the seventies
38	as his farewell appearances around into his seventies
39	as his farewell appearances again on into his seventies
40	has his farewell appearances on into the seventies
41	as his farewell appearances langer on into his seventies
42	as his farewell appearances like iran in his seventies
43	has his farewell appearances drawn into his seventies
44	as his farewell appearances on in his seventies
45	as his farewell appearances go on into his seventies
46	as his farewell appearances i go on into the seventies
45	as his farewell appearances in iran in the seventies
48	as his farewell appearances iran in his seventies
49	as his farewell appearances stagger on in his seventies
50	as his farewell appearances in on into the seventies

Example of ASR N-best List

Source: European Parliament Plenary Sessions

1	as his farewell appearances	like iran	into his seventies	26	as his farewell appearances	tag and on	into the seventies
2	as his farewell appearances	iran	into his seventies	27	as his farewell appearances	like iran	and the seventies
3	as his farewell appearances	on	into his seventies	29	as his farewell appearances	here on	in the seventies
4	as his farewell appearances	on	into the seventies	30	as his farewell appearance	as iran	into his seventies
5	as his farewell appearances	drawn	into his seventies	31	as his farewell appearances	like and on	into the seventies
6	as his farewell appearances	like iran	into the seventies	32	as his farewell appearances	that iran	into his seventies
7	as his farewell appearances	iran	into the seventies	33	as his farewell appearances	i go on	into his seventies
8	as his farewell appearances	drawn	into the seventies	34	as his farewell appearances	stagger on	into the seventies
9	as his farewell appearances	and on	into his seventies	35	as his farewell appearances	in on	into his seventies
10	has his farewell appearances	like iran	into his seventies	36	as his farewell appearances	by iran	into his seventies
11	as his farewell appearances	like enron	into his seventies	37	as his farewell appearances	here on	into the seventies
12	as his farewell appearances	like iran	in the seventies	38	as his farewell appearances	around	into his seventies
13	has his farewell appearances	iran	into his seventies	39	as his farewell appearances	again on	into his seventies
15	as his farewell appearances	iran	in the seventies	40	has his farewell appearances	on	into the seventies
16	as his farewell appearances	and on	into the seventies	41	as his farewell appearances	langer on	into his seventies
17	as his farewell appearances	tag and on	into his seventies	42	as his farewell appearances	like iran	in his seventies
18	as his farewell appearances	gone	into his seventies	43	has his farewell appearances	drawn	into his seventies
19	as his farewell appearances	on	in the seventies	44	as his farewell appearances	on	in his seventies
20	as his farewell appearances	in iran	into his seventies	45	as his farewell appearances	go on	into his seventies
21	as his farewell appearances	like and on	into his seventies	46	as his farewell appearances	i go on	into the seventies
22	as his farewell appearances	stagger on	in the seventies	45	as his farewell appearances	in iran	in the seventies
23	as his farewell appearances	stagger on	into his seventies	48	as his farewell appearances	iran	in his seventies
24	as his farewell appearances	here on	into his seventies	49	as his farewell appearances	stagger on	in his seventies
25	has his farewell appearances	on	into his seventies	50	as his farewell appearances	in on	into the seventies

Experiments with N-Best Decoding

Translation results on the BTEC Italian-English task [Quan et al., 2005]:

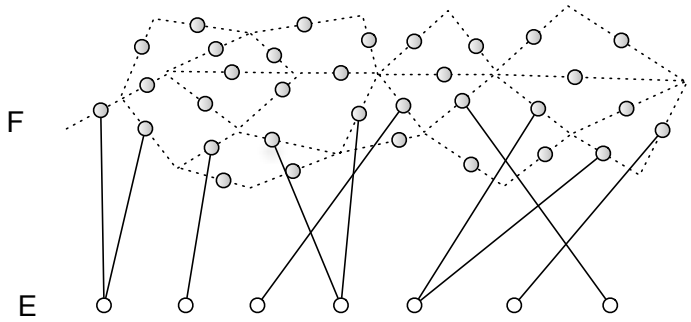
system	BLEU	95% conf. interval
1-Best	40.02	38.88 - 41.18
100-Best	41.22	40.03 - 42.42

- Step 1. DP beam search decoder with log-linear model
  – applied to each 100-best produces each time 100-best translations
- Step 2. Re-scoring module with log-linear model
  – applied to all 100x100 translations
  – features functions include source language model and acoustic model scores

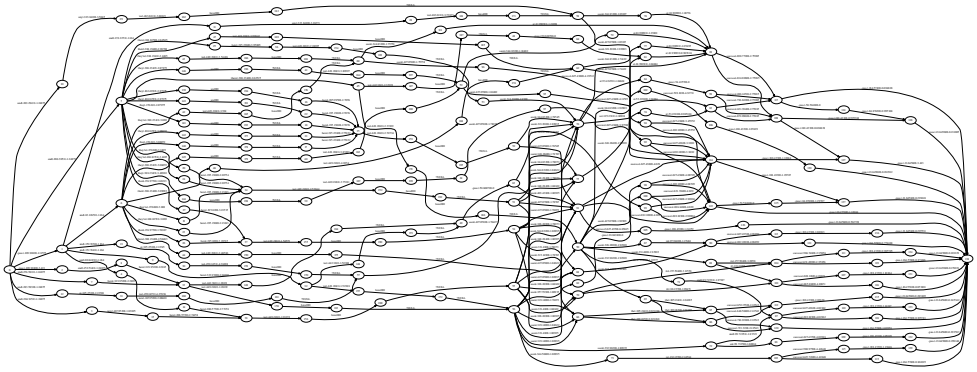
Word-Graph Decoding

$\mathcal{H}(\mathbf{x})$ include all paths in the ASR WG [Matusov et al., 2006, Saleem et al., 2004]

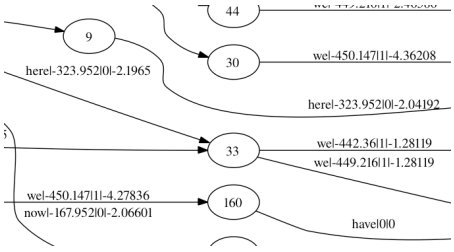
- **Pros: easy to integrate in the SFST framework** [Mathias and Byrne, 2006]
  - permits to explore a huge set of ASR hypotheses
  - efficient decoders for quasi-monotonic search [Zhou et al., 2007]
- **Cons: search space grows very quickly**:
  - difficult to apply for large domain tasks
  - long word re-ordering severely impacts on complexity



A Real Word-Graph



ZOOM



Experiments with Word-Graph Decoding

BTEC Italian-English task [Matusov et al., 2006] comparing:

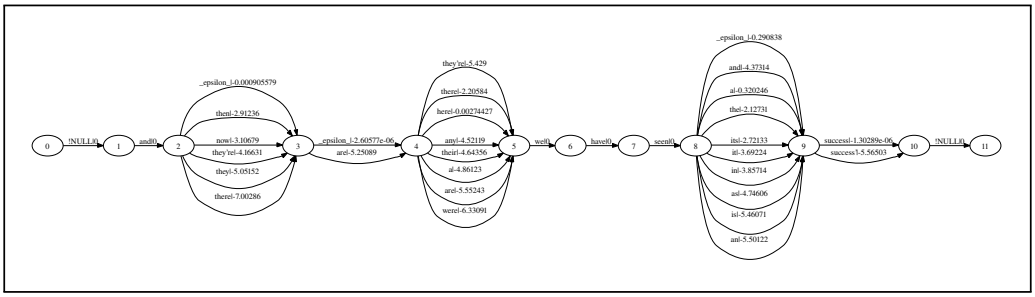
- FST phrase-based approach with joint probabilities (FST)
- log-linear phrase-based approach (monotone search over WG) (PBT).

Approach	Interface	WER	PER	BLEU
FST	single best	33.4	29.1	52.7
	word graph + acoustic scores	31.6	27.6	54.3
PBT	single best	32.4	27.2	55.4
	word graph	31.9	28.0	54.7
	+ acoustic and LM scores	30.6	26.6	56.2
	+ optimized exponents	29.8	25.8	57.7

Confusion Network

Confusion networks (Mangu, 1999) approximate WGs by linear networks, s.t.:

- arcs are labeled with words or with the *empty word* (ϵ-word)
- arcs are weighted with word *posterior probabilities*
- paths are a superset of those in the word graph

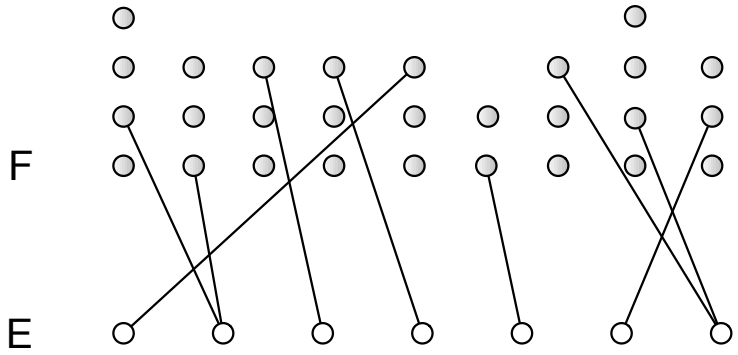


CNs can be conveniently represented as a sequences of *columns* of different depth.

Confusion Network Decoding

$\mathcal{H}(\mathbf{x})$ includes all paths of the ASR confusion network [Bertoldi and Federico, 2005]

- **Pros: easy to integrated with log-linear translation models**
  - permits to explore a huge search space
  - **very efficient decoder** [Bertoldi et al., 2007]
- **Cons: confusion network contain less information than WGs**:
  - *spurious paths* are introduced (however with low probability)
  - posterior scores on arcs are approximations (*independence assumption*)



Example of Confusion Network (1)

Source: European Parliament Plenary Sessions

as	his	farewell	appearances	ϵ	ϵ	iran	into	his	seventies
has	its	fellow	appearance	like	and	on	in	the	seventy's
and	is			stagger	go	drawn	and	their	
				and	i	enron	until		
				tag	you	gone			
				in		around			
				i		non			
				going					
				again					
				...					

N-BEST:
1 as his farewell appearances like iran into his seventies
2 as his farewell appearances iran into his seventies
3 as his farewell appearances on into his seventies
4 as his farewell appearances on into the seventies
5 as his farewell appearances drawn into his seventies

Example of Confusion Network (2)

Source: European Parliament Plenary Sessions

as	his	farewell	appearances	ϵ	ϵ	iran	into	his	seventies
has	its	fellow	appearance	like	and	on	in	the	seventy's
and	is			stagger	go	drawn	and	their	
				and	i	enron	until		
				tag	you	gone			
				in		around			
				i		non			
				going					
				again					
				...					

N-BEST:
1. as his farewell appearances like iran into his seventies
2. as his farewell appearances iran into his seventies
3. as his farewell appearances on into his seventies
4. as his farewell appearances on into the seventies
5. as his farewell appearances drawn into his seventies

Example of Confusion Network (3)

Source: European Parliament Plenary Sessions

as	his	farewell	appearances	ϵ	ϵ	iran	into	his	seventies
has	its	fellow	appearance	like	and	on	in	the	seventy's
and	is			stagger	go	drawn	and	their	
				and	i	enron	until		
				tag	you	gone			
				in		around			
				i		non			
				going					
				again					
				...					

N-BEST:
1. as his farewell appearances like iran into his seventies
2. as his farewell appearances iran into his seventies
3. as his farewell appearances on into his seventies
4. as his farewell appearances on into the seventies
5. as his farewell appearances drawn into his seventies

Example of Confusion Network (4)

Source: European Parliament Plenary Sessions

as	his	farewell	appearances	ϵ	ϵ	iran	into	his	seventies
has	its	fellow	appearance	like	and	on	in	the	seventy's
and	is			stagger	go	drawn	and	their	
				and	i	enron	until		
				tag	you	gone			
				in		around			
				i		non			
				going					
				again					
				...					

N-BEST:
1 as his farewell appearances like iran into his seventies
2 as his farewell appearances iran into his seventies
3 as his farewell appearances on into his seventies
4 as his farewell appearances on into the seventies
5 as his farewell appearances drawn into his seventies

Confusion Network Translation: Search

Extension of basic phrase-based decoding step:

- **cover** some not yet covered consecutive columns (*span*)
- **retrieve** phrase-translations for all paths inside the columns
- **compute** translation, distortion and target language models

Example. Coverage set: 01110... Path: *cancello d'*

0	1	1	1	0	. ..
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		all' 0.005		
			l' 0.002		
			ϵ 0.001		

Confusion Network Translation: Search

Computational issues:

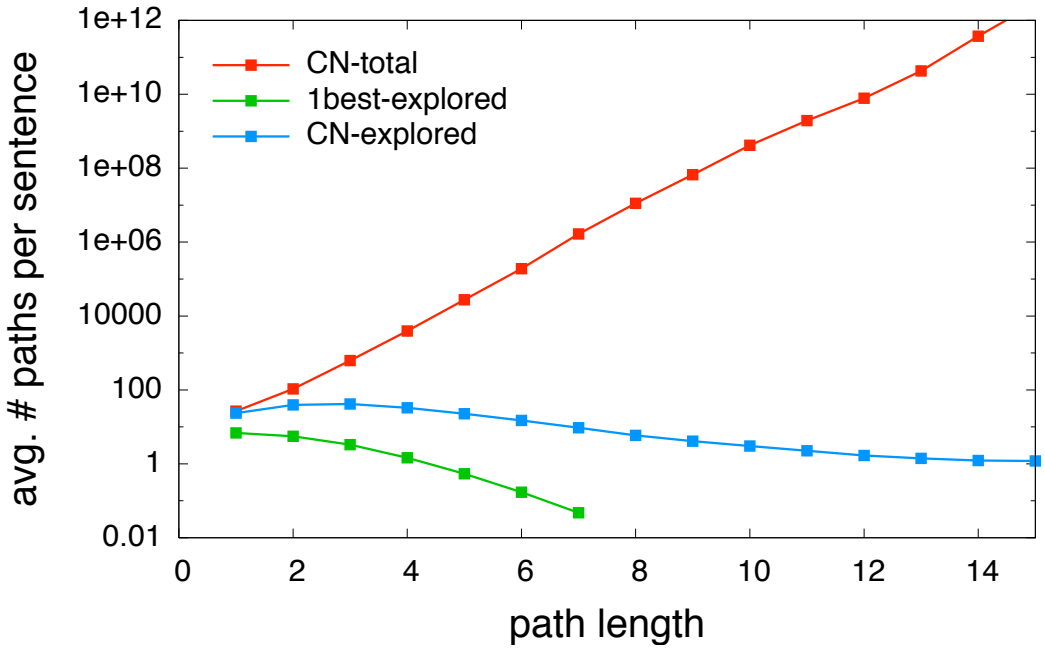
- Number of paths grows **exponentially** with span length
- Implies look-up of translations for a huge number of source phrases
- Factored models require considering joint translation over all factors (tuples):
  – cartesian product of all translations of each single factor

Solutions implemented into Moses

- Source entries of the phrase-table are stored with *prefix-trees*
- Translations of all possible coverage sets are *pre-fetched from disk*
- Efficiency achieved by *incrementally pre-fetching* over the span length
- Phrase translations over all factors are extracted independently, then translation tuples are generated and pruned by adding a factor each time

Once translation tuples are generated, usual decoding applies.

Confusion Network Translation: Efficiency



Confusion Network Translation: Experiments

Translation results [%] on the EPPS Spanish-English task: comparison of the single-best, N-best, and confusion network interfaces [Bertoldi et al., 2007].

Spanish-English Task

Input		Output			
type	WER	BLEU	NIST	WER	PER
verbatim	0.0	48.00	9.864	40.96	31.19
wg-oracle	7.48	44.68	9.507	43.75	33.55
cn-oracle	8.45	44.12	9.356	44.95	34.37
1-best	22.41	37.57	8.590	50.01	39.24
cons-dec	23.30	36.98	8.550	49.98	39.17
cn	8.45	39.17	8.716	49.52	38.64
10-best	17.12	38.71	8.670	49.29	38.74
100-best	13.68	38.95	8.695	49.19	38.69

Part 3: The TC-STAR Project

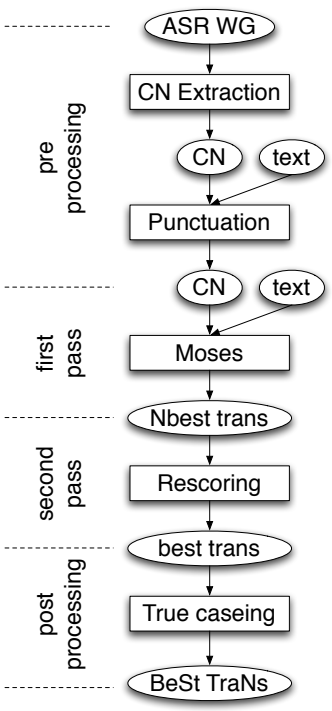
Research Challenges in TC-STAR

Spoken language translation of **political speeches** (unrestricted domain)

- *Tasks*:
  - EU Parliament Plenary Sessions: Spanish-English
  - Spanish Parliament: Spanish-to-English
- *Types of input*:
  - Automatic speech recognition output (ASR)
  - Verbatim transcription (VBT): literal transcript of a speech
  - Final text editions (FTE): polished texts

Spoken language translation of **broadcast news** (unrestricted domain)

- *Task*: Voice of America: Mandarin-to-Chinese
- *Types of input*:
  - Automatic speech recognition output (ASR)
  - Verbatim transcription (VBT): literal transcript of a speech



FBK Translation System

- *Processing of full confusion networks*
- *Punctuating confusion networks*
- *Moses: efficient CN decoder*
- *IRSTLM: to handle huge LMs*
- *Re-scoring with new feature functions*

Moses Toolkit for Statistical MT

- Developed during **JHU Summer Workshop 2006**
  - U. Edinburgh, ITC-irst Trento, RWTH Aachen, U. Maryland, MIT, Charles University Prague
  - open source under Lesser GPL
  - available for Linux, Windows and Mac OS
  - www.statmt.org/moses
- **Main features**:
  - *translation of both text and CN inputs*
  - exploitation of more Language Models
  - lexicalized distortion model (only for text input, optional)
  - incremental pre-fetching of translation options from disk
  - *handling of huge LMs* (up to Giga words)
  - *on-demand and on-disk access to LMs* and LexMs
  - factored translation model (surface forms, lemma, POS, word classes, ...)

Punctuating Confusion Networks

Confusion network without punctuation

i.9 hi.1	cannot.8 can.1 ϵ.1	ϵ.7 not.3	say.6 said.2 say.1 ϵ.1	ϵ.7 any.3	anything.8 thing.1 things.1	at.9 ϵ.1	this.8 these.1 those.1	point.7 points.1 ϵ.1 pint.1	are1	there.8 the.1 their.1	ϵ.8 a.1 air.1	any.7 new.1 a.1 ϵ.1	comments.7 comment.2 commit.1

Consensus decoding

i cannot say anything at this point are there any comments

Punctuating confusion network

i1	cannot1	say1	anything1	ϵ.9 ·.1	at1	this1	point1	·.7 ϵ.2 ?.1	are1	there1	any1	comments1	?.6 ϵ.3 ·.1

Punctuated confusion network

i.9 hi.1	cannot.8 can.1 ϵ.1	ϵ.7 not.3	say.6 said.2 say.1 ϵ.1	ϵ.7 any.3	anything.8 thing.1 things.1	ϵ.9 ·.1	at.9 ϵ.1	this.8 these.1 those.1	point.7 points.1 ϵ.1 pint.1	·.7 ϵ.2 ?.1	are1	there.8 the.1 their.1	ϵ.8 a.1 air.1	any.7 new.1 a.1 ϵ.1	comments.7 comment.2 commit.1	?.6 ϵ.3 ·.1

Punctuating Confusion Networks: Results

- ASR 1-best output vs. confusion network
- 1-best punctuation vs. punctuating CN (from 1K-best)

Spanish-English EPPS Eval06					
ASR type	punctuation	BLEU	NIST	WER	PER
1-best	1-best	35.62	8.37	57.15	44.56
	CN	36.01	8.41	56.78	44.39
CN	1-best	36.22	8.46	56.39	44.37
	CN	36.45	8.49	56.17	44.19

Chinese-English System

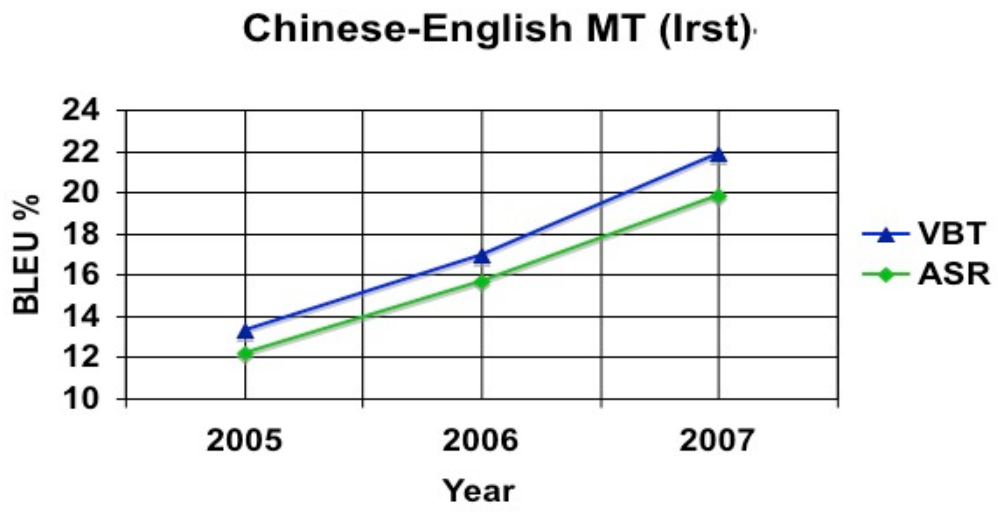
- **Target Language Models**
  - 3 LMs: target part of parallel data + GigaWord + DevSets
  - 2G running words (4.5M different words)
  - 300M 5-grams (singletons pruned for GigaWord)
- **Phrase Table**
  - 90M English running words
  - 38M phrase pairs of maximum length 7
- **Lexicalized reordering model**
  - conditioned on both the source and target phrases

Chinese-English System

- **Miscellanea & ad hoc tricks**
  - Re-scoring with 6-gram LM of target side
  - Punctuation removed from translation
  - Digits transformed into text
  - Pinyin transliteration of untranslated Mandarin words

- **Submissions**
  - primary condition
  - VHT and ASR rover
  - no CN decoder

Results Overview



2007: second top system of seven participants

Chinese-English Translations

Human	Primakov made the above announcement after a meeting held by government on Saturday.
MT 05	Primakov on Saturday at a time when the Government after the meeting as well as making the announcement.
MT 06	Primakov on Saturday held a meeting made following the government announced.
MT 07	Primakov on Saturday at the first meeting after the government made the above announcement.

Chinese-English Translations (2)

Human	On Saturday Chinese TV station broadcasted a long documentary describing the early life of Mao Zedong in memory of his birthday anniversary
MT 05	China T.v. Saturday broadcast by a minute and led early in life long documentaries in memory of Mao Zedong's Birthday
MT 06	Chinese television Saturday broadcast a descriptive of Mao Zedong early life long documentary commemorating Mao Zedong's birthday
MT 07	Chinese television broadcast on Saturday a long documentary early life of Mao Mao Zedong Memorial birthday

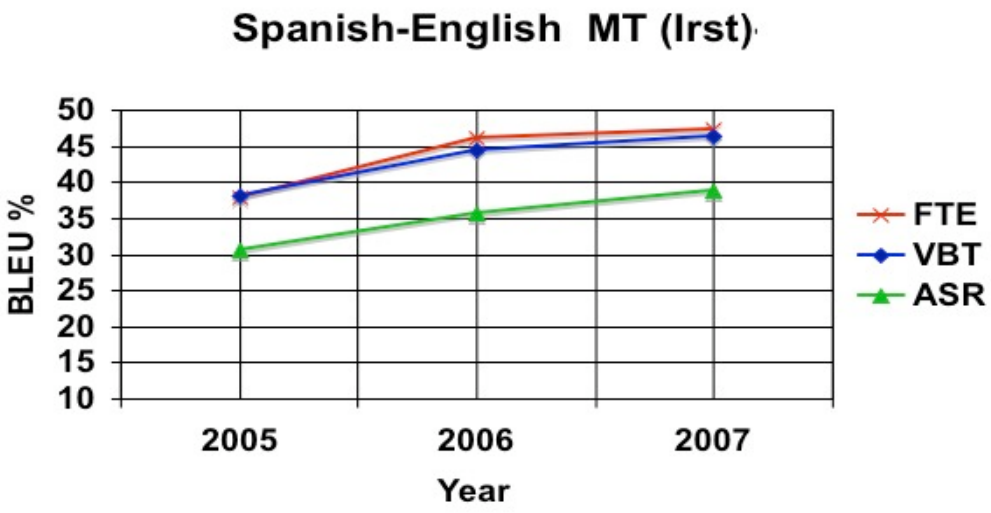
Spanish-English Systems

- **English Language Models**
  - 4 resources: EPPS + GigaWord + DevSets + target part of other bilingual corpora (EU-bulletin, JRC-Acquis, UN)
  - 2G running words (4.5M different words)
  - 350M 5-grams (singletons pruned for GigaWord)
- **Spanish Language Models**
  - 4 resources: EPPS + GigaWord + DevSets + target part of other bilingual corpora (EU-bulletin, JRC-Acquis, UN)
  - 780M running words (1.7M different words)
  - 140M 5-grams (singletons pruned for GigaWord)
- **Phrase Table**
  - 37M English and 36M Spanish running words
  - 83M phrase pairs of maximum length 8

English-Spanish Systems

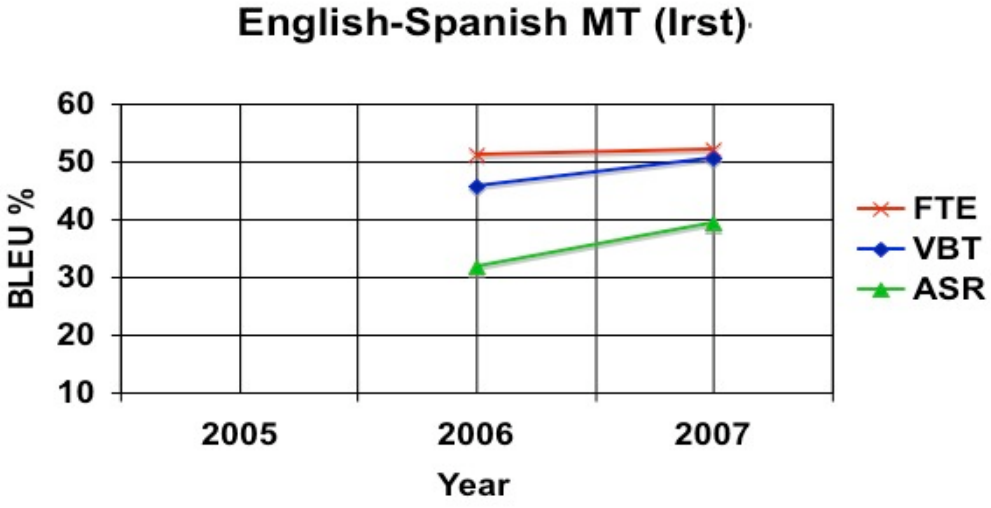
- **Submissions: primary condition**
  - text decoder: FTE, VHT and ASR rover
  - CN decoder: re-punctuated ASR rover, CN and consensus decoding from LIMSI lattices
- **Miscellanea**
  - Spanish re-scoring 6-gram LM estimated on EPPS + other corpora
  - punctuation insertion for ASR condition only
  - weight optimization on VHT condition only

Progress in SLT



2007: top system on ASR (case insensitive output)

Progress in SLT



2007: top system on ASR (case insensitive output)

Spanish-English Translations

Human	What is happening with the adopted plan to prevent and combat trafficking in human beings?
MT 05	What is happening with the plan was adopted for preventing and combating trafficking in human beings ?
MT 06	What is happening with the plan adopted for preventing and combating the trafficking of human beings ?
MT 07	What is happening with the plan adopted to prevent and combat the trafficking of human beings ?

Spanish-English Translations (2)

HT	On behalf of the European Parliament I would like to join their families in their grief.
MT 05	On behalf of the European Parliament would like to join the pain of their families .
MT 06	on behalf of the European Parliament would like to join the pain of their families .
MT 07	On behalf of the European Parliament I would like to join the pain of their families .

References

[Bangalore and Riccardi, 2003] Bangalore, S. and Riccardi, G. (2003). Stochastic finite-state models for spoken language machine translation. *Machine Translation*, 17(3):165–184.

[Bertoldi and Federico, 2005] Bertoldi, N. and Federico, M. (2005). A new decoder for spoken language translation based on confusion networks. In *Proceedings of the IEEE ASRU Workshop*, pages 86–91, San Juan, Puerto Rico.

[Bertoldi et al., 2007] Bertoldi, N., Zens, R., and Federico, M. (2007). Speech translation by confusion network decoding. In *Proceedings of ICASSP*, pages 1297–1300, Honolulu, HA.

[Casacuberta and Vidal, 2004] Casacuberta, F. and Vidal, E. (2004). Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225.

[Casacuberta and Vidal, 2007] Casacuberta, F. and Vidal, E. (2007). Learning finite-state models for machine translation. *Machine Learning*, 66(1):69–91.

[Federico and Bertoldi, 2005] Federico, M. and Bertoldi, N. (2005). A word-to-phrase statistical translation model. *ACM Transactions on Speech and Language Processing (TSLP)*, 2(2):1–24.

[Knight and Al-Onaizan, 1998] Knight, K. and Al-Onaizan, Y. (1998). Translation with finite-state devices. In *Proceedings of the AMTA Conference*, pages 421–437, Langhorne, PA.

[Koehn et al., 2003] Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of HLT/NAACL 2003*, pages 127–133, Edmonton, Canada.

[Kumar et al., 2006] Kumar, S., Deng, Y., and Byrne, W. (2006). A weighted finite state transducer translation template model for statistical machine translation. *Journal of Natural Language Engineering,*, 12(1):35–75.

[Mariño et al., 2006] Mariño, J. B., Banchs, R. E., Crego, J. M., de Gispert, A., Lambert, P., Fonollosa, J. A. R., and Costa-Jussà, M. R. (2006). N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

[Mathias and Byrne, 2006] Mathias, L. and Byrne, W. (2006). Statistical phrase-based speech translation. In *Proceedings of ICASSP*, pages 561–564, Toulouse, France.

[Matusov et al., 2006] Matusov, E., Kanthak, S., and Ney, H. (2006). Integrating speech recognition and machine translation: Where do we stand? In *Proceedings of ICASSP*, pages 1217–1220, Toulouse, France.

[Och and Ney, 2002] Och, F. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL*, pages 295–302, Philadelhpia, PA.

[Och and Ney, 2003] Och, F. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

[Quan et al., 2005] Quan, V., Federico, M., and Cettolo, M. (2005). Integrated N-Best Re-Ranking for Spoken Language Translation. ISCA.

[Saleem et al., 2004] Saleem, S., Jou, S.-C., Vogel, S., and Schultz, T. (2004). Using word lattice information for a tighter coupling in speech translation systems. In *Proceedings of ICASSP*, pages 765–768, Jeju Island, South Korea.

[Tillmann and Ney, 2003] Tillmann, C. and Ney, H. (2003). Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133.

[Zhang et al., 2004] Zhang, R., Kikui, G., Yamamoto, H., Watanabe, T., Soong, F., and Lo, W. K. (2004). A unified approach in speech-to-speech tanslation: integrating features of speech recognition and machine translation. In *Proceedings of COLING*, pages 1168–1174, Geneva, Switzerland.

[Zhou et al., 2007] Zhou, B., Besacier, L., and Gao, Y. (2007). On efficient coupling of ASR and SMT for speech translation. In *Proceedings of ICASSP*, pages 101–104, Honolulu, HA.