MEMT: Alignment-based MT System Combination with Linguistic and Statistical Features

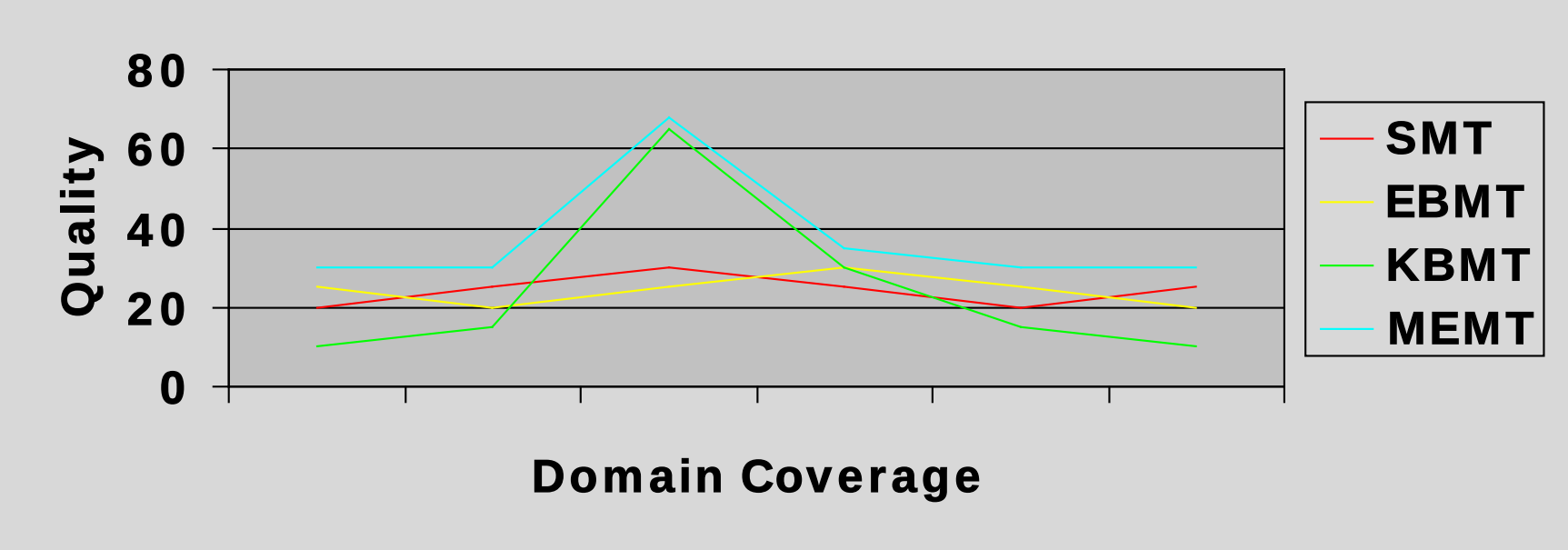
Alon Lavie

Language Technologies Institute
Carnegie Mellon University
19 November 2011

Joint work with:
Kenneth Heafield, Greg Hanneman and Michael Denkowski

Carnegie Mellon

MT System Combination



MT System Combination

- **Idea:** apply several MT engines to each input in parallel and combine their output translations
- **Goal:** leverage the strengths and diversity of different MT engines to generate an improved translation system
- Particularly useful in **assimilation scenarios** where input is uncontrolled and diverse in domain, genre, style or other characteristics
- Can result in significant gains in translation quality

Main Issues and Challenges

- Selecting the "best" output among the engines (on a sentence-by-sentence basis), or generating a deeper combination of the translation output from the original engines?
  - Selecting between engines is easier
  - Synthetically combining them can potentially produce greater improvements in translation quality
- What information is available from each of the MT engines?
- How do we determine if a synthetically combined translation is better than the originals?

Consensus Network Approach

- Main Ideas:
  - Collapse the collection of linear strings of multiple translations into a minimal consensus network ("sausage" graph) that represents how they align
    - How are the system outputs aligned?
  - One translation acts as the "back-bone" and determines the main ordering of words
  - **Decode:** find the path through the network that has the best score
- Main Weaknesses:
  - Search is limited to selecting among aligned alternatives along the back-bone
  - Prone to dropping words due to "epsilon" arcs
  - Long distance alternations can result in repetitions

Consensus Network Example

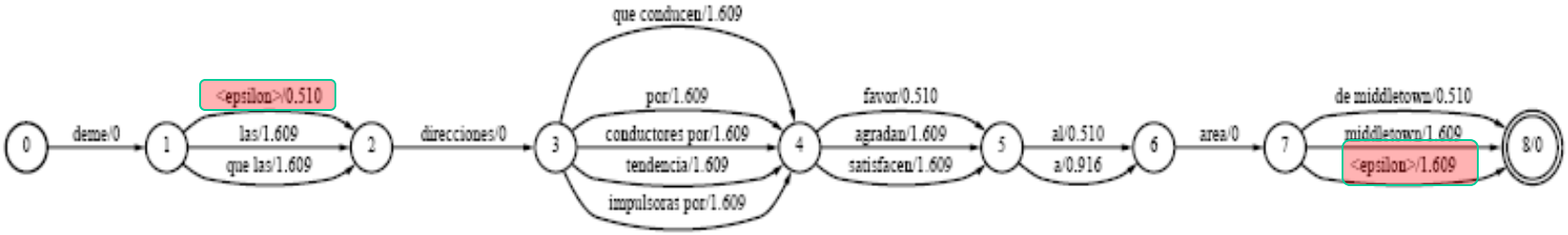


Fig. 4. Lattice representation of the result of the multiple alignment. The weights on the arcs are negative logarithm of the probability that word.

CMU's Alignment-based Multi-Engine System Combination

- Works with any MT engines
  - Assumes original MT systems are "black-boxes" – no internal information other than the translations themselves
- Explores broader search spaces than other MT system combination approaches using linguistically-based and statistical features
- Achieves state-of-the-art performance in research evaluations over past couple of years
- Developed over last five years under research funding from several government grants (DARPA, DoD and NSF)

Alignment-based MEMT

Two Stage Approach:

1. **Align:** Identify and align equivalent words and phrases across the translations provided by the engines
2. **Decode:** search the space of synthetic combinations of words/phrases and select the highest scoring combined translation

Example:

1. announced afghan authorities on saturday reconstituted four intergovernmental committees
2. The Afghan authorities on Saturday the formation of the four committees of government

Alignment-based MEMT

Two Stage Approach:

1. **Align:** Identify and align equivalent words and phrases across the translations provided by the engines
2. **Decode:** search the space of synthetic combinations of words/phrases and select the highest scoring combined translation

Example:

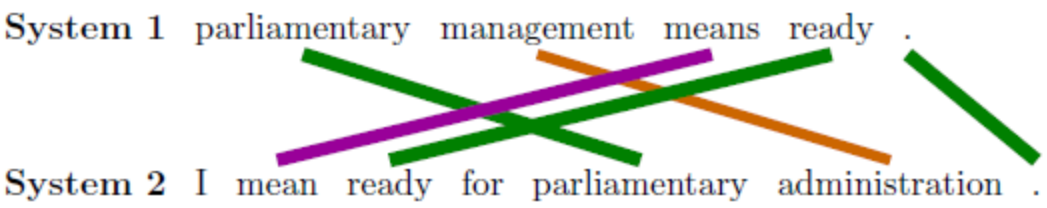
1. announced afghan authorities on saturday reconstituted four intergovernmental committees
2. The Afghan authorities on Saturday the formation of the four committees of government

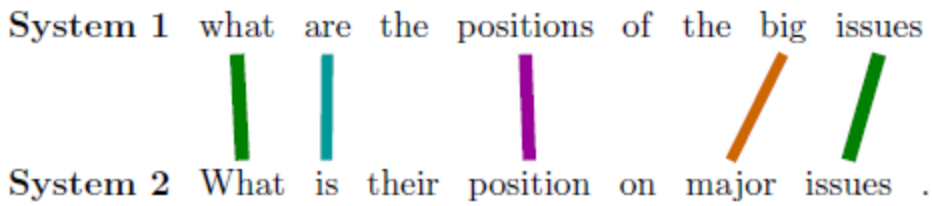
MEMT: the afghan authorities announced on Saturday the formation of four intergovernmental committees

The String Alignment Matcher

- Developed as a component in our METEOR Automated MT Evaluation system
- Originally word-based, extended to phrasal matches
- Finds maximal alignment match with minimal “crossing branches” (reordering)
- Allows alignment of:
  - Identical words
  - Morphological variants of words (using stemming)
  - Synonymous words (based on WordNet synsets)
  - Single and multi-word Paraphrases (based on statistically-learned paraphrase tables)
- **Implementation:** efficient search algorithm for best scoring weighed string match

The String Alignment Matcher Examples:

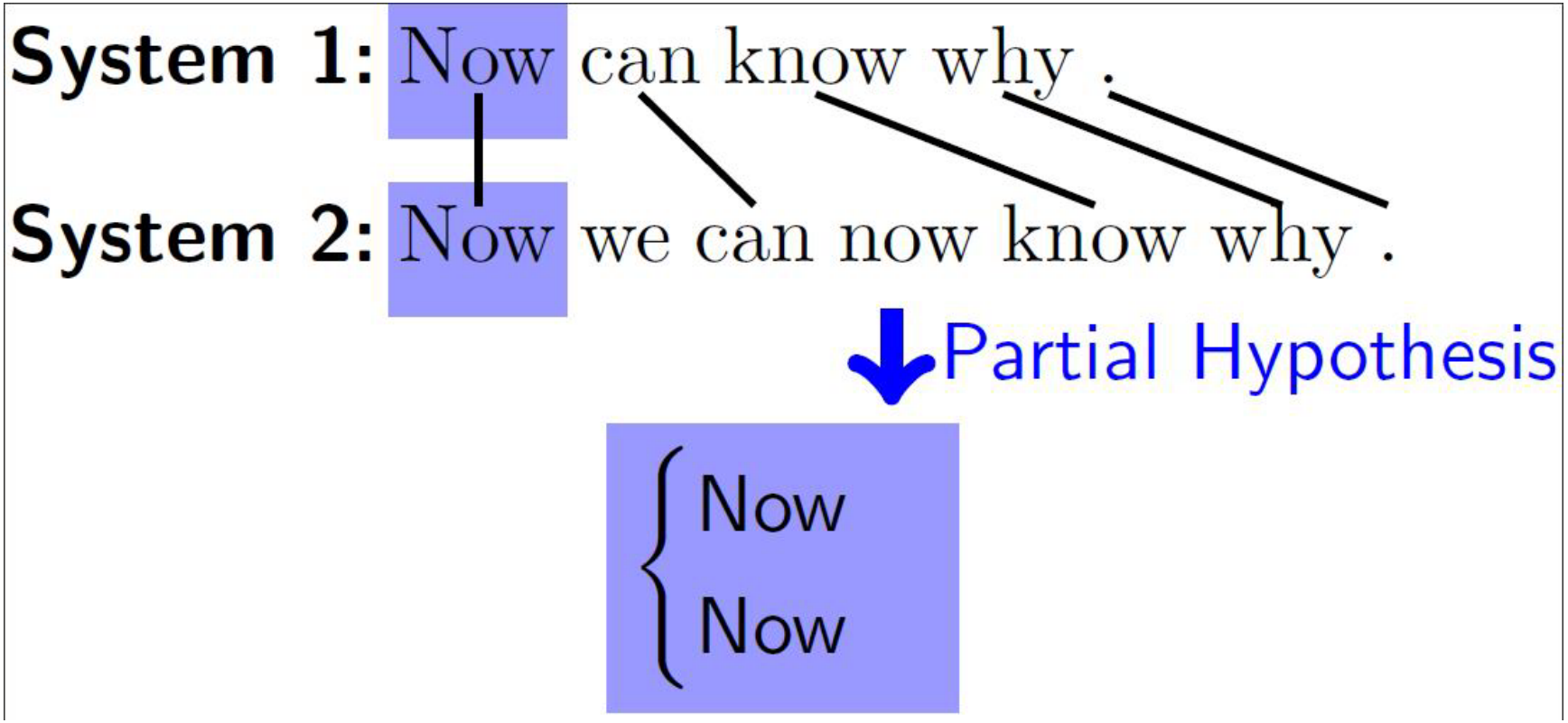




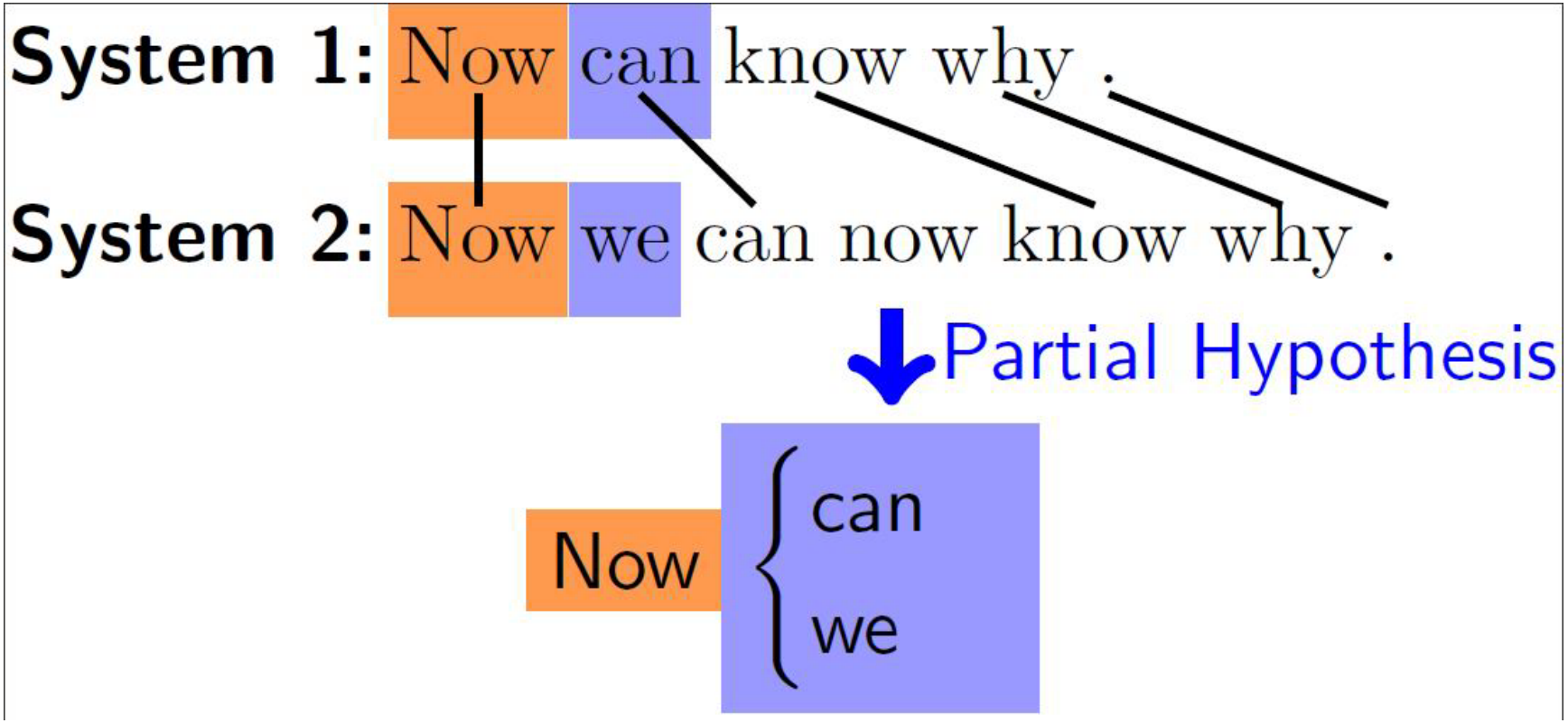
The MEMT Decoder Algorithm

- **Search-space** of system combination hypotheses implicitly defined by the initial alignment stage, and partially explored
  - Search-space is controlled by **linguistic similarity features**
- Algorithm builds collections of partial hypotheses of increasing length
- Partial hypotheses are extended by selecting the "next available" word from one of the original systems
- Extending a partial hypothesis with a word marks the word as "used" and marks its aligned words as also "used"
- Partial hypotheses are scored and ranked
- Pruning and re-combination for efficiency
- Hypothesis can end if any original system proposes an end of sentence as next word

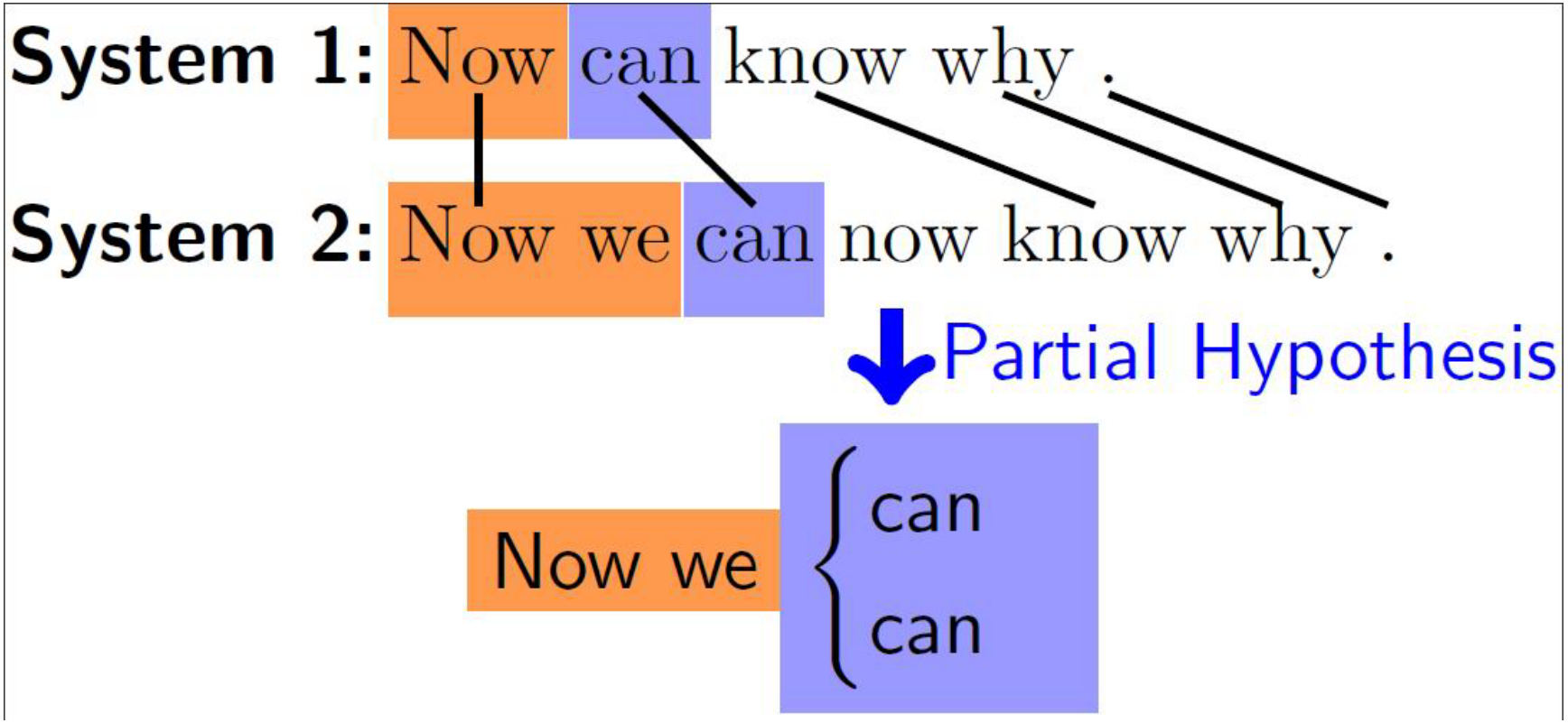
Decoding Example



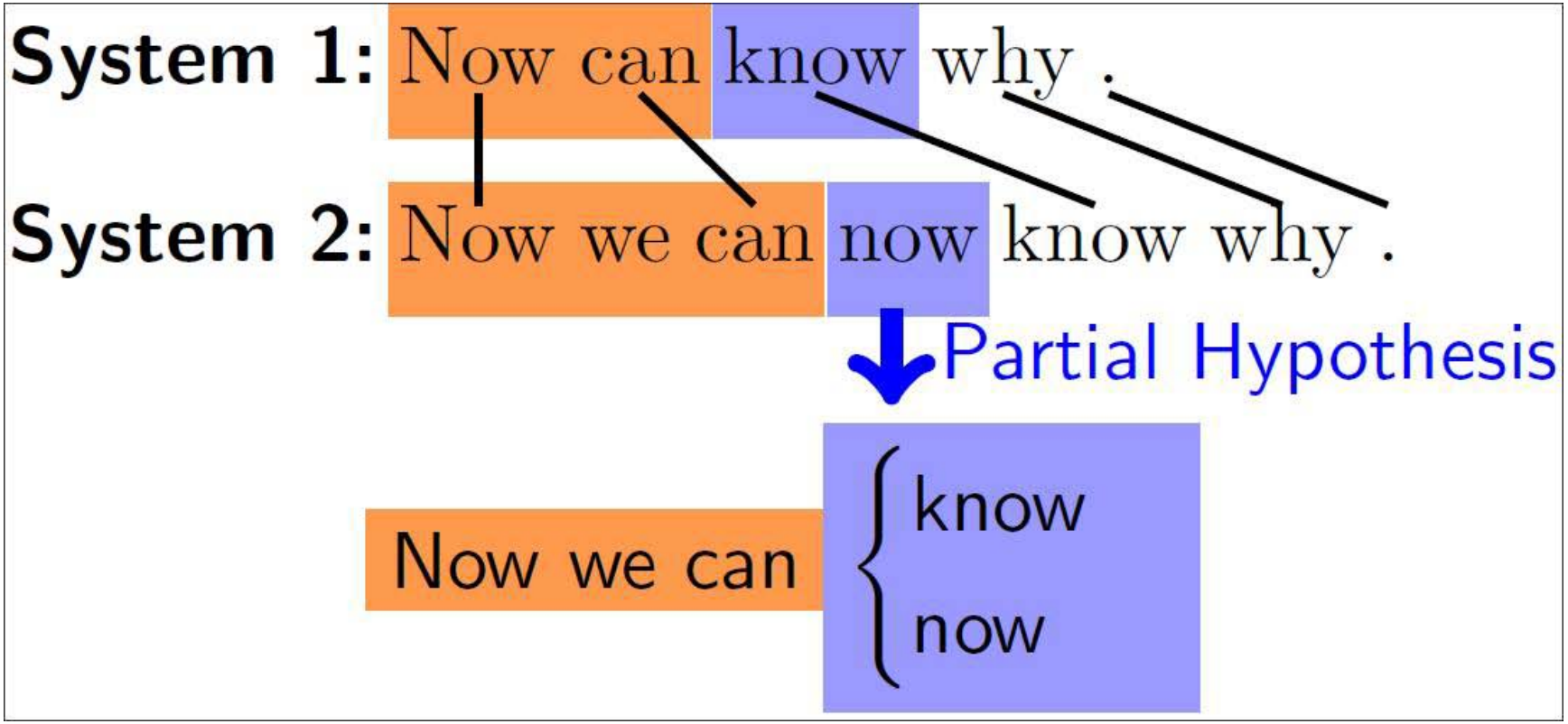
Decoding Example



Decoding Example



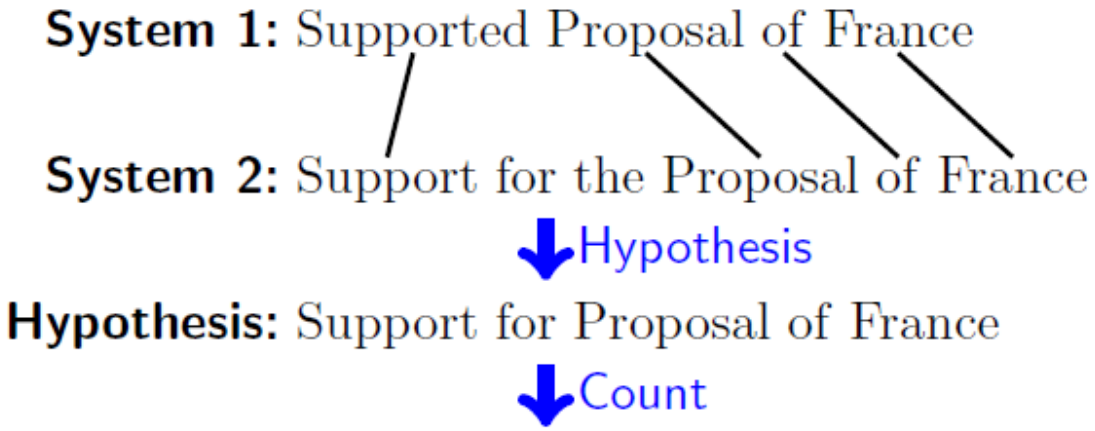
Decoding Example



Scoring MEMT Hypotheses

- **Features:**
  - N-gram Language Model score based on filtered large-scale target language LM
  - N-gram support features with n-grams matches from the original systems (unigrams to 4-grams)
  - Length
- **Scoring:**
  - Weighed Log-linear feature combination tuned on development set
  - Weights are tuned using MERT on a held-out tuning set

N-gram Match Support Features



	Unigram	Bigram	Trigram	Quadgram
System 1	4	2	1	0
System 2	5	3	1	0

Hyper-Parameters

- Selecting among the various MT systems available for combination
  - Combine all or just a subset?
  - Criteria for selection: metric scores, diversity of approach, other…
- Internal Hyper-settings:
  - "Horizon": when to drop lingering words
  - N-gram match support features: per individual system or aggregate across systems?
- Highly efficient implementation allows executing exhaustive collection of experiments with different hyper-parameter settings on distributed parallel high-computing clusters

Recent Performance Results NIST-2009 and WMT-2009

Source	Top	Gain
Arabic	58.55	+6.67
Czech	21.98	+0.80
French	31.56	+0.42
German	23.88	+2.57
Hungarian	13.84	+1.09
Spanish	28.79	+0.10
Urdu	34.72	+1.84

Table: Post-evaluation uncased BLEU gains on NIST and WMT tasks.

Recent Performance Results WMT-2010

French-English
589–716 judgments per combo

System	≥others
RWTH-COMBO ●	0.77
CMU-HYP-COMBO ●	0.77
DCU-COMBO ●	0.72
LIUM ★	0.71
CMU-HEA-COMBO ●	0.70
UPV-COMBO ●	0.68
NRC	0.66
CAMBRIDGE	0.66
UEDIN ★	0.65
LIMSI ★	0.65
JHU-COMBO	0.65
RALI	0.65
LIUM-COMBO	0.64
BBN-COMBO	0.64
RWTH	0.55

English-French
740–829 judgments per combo

System	≥others
RWTH-COMBO ●	0.75
CMU-HEA-COMBO ●	0.74
UEDIN	0.70
KOC-COMBO ●	0.68
UPV-COMBO	0.66
RALI ★	0.66
LIMSI	0.66
RWTH	0.63
CAMBRIDGE	0.63

Recent Performance Results WMT-2010

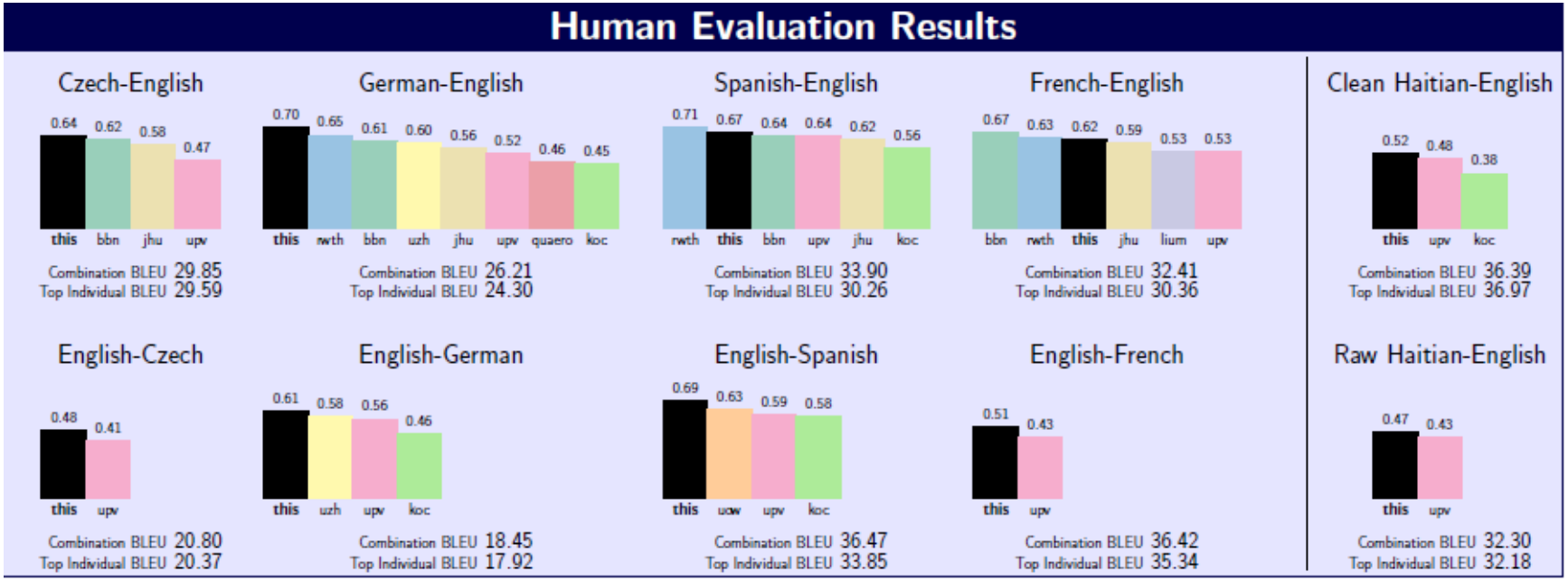
Spanish-English
1385–1535 judgments per combo

System	≥others
UEDIN ⋆	0.69
CMU-HEA-COMBO •	0.66
UPV-COMBO •	0.66
BBN-COMBO	0.62
JHU-COMBO	0.55
UPC	0.51

English-Spanish
516–673 judgments per combo

System	≥others
CMU-HEA-COMBO •	0.68
KOC-COMBO	0.62
UEDIN ⋆	0.61
UPV-COMBO	0.60
RWTH-COMBO	0.59
DFKI ⋆	0.55
JHU	0.55
UPV	0.55
CAMBRIDGE ⋆	0.54
UPV-NNLM ⋆	0.54

Recent Performance Results WMT-2011



Smoothing MERT in SMT [Cettolo, Bertoldi and Federico 2011]

- Interesting application of MT system combination to overcome instability of MERT optimization in SMT
  - Perform MERT multiple times
  - Use the CMU MEMT system to combine the different instances of **the same MT system**

en–fi	BLEU%	stdev	[min,max]
optSample	35.95	0.080	[35.83,36.07]
avg6	35.97	0.023	[35.93,36.01]
sysComb6	36.34	0.106	[36.21,36.50]

el–fr	BLEU%	stdev	[min,max]
optSample	58.22	0.104	[58.01,58.33]
avg6	58.09	0.043	[58.02,58.15]
sysComb6	58.92	0.114	[58.71,59.08]

Table 4: Results for the ACQUIS task on the test set.

CMU MEMT System is Open Source

- http://kheafield.com/code/memt/
- Open Source, LGPL license
- Freely available for research and commercial use

Current and Future Work

- Incorporation of multi-word paraphrase matches into the decoding algorithm
- Improved search-space exploration
  - Linguistically-motivated constraints?
- Additional scoring features
  - Linguistically-motivated features?
- Second-pass MBR-decoding over n-best lists
- Multi-Engine **Human** Translation