MT System Combination

11-731
Machine Translation
Alon Lavie
April 9, 2015

With acknowledged contributions from Silja Hildebrand and Kenneth Heafield

Goals and Challenges

- Different MT systems have different strengths and weaknesses
  - Different approaches: Phrase-based, Hierarchical, Syntax-based, RBMT, EBMT
  - Different domains, training data, tuning data
- Scientific Challenge:
  - How to combine the output of multiple MT engines into a selected output that outperforms the originals in translation quality?
- Selecting the best output on a sentence-by-sentence basis (classification), or a more synthetic combination?
- Range of approaches to address the problem
- Can result in very significant gains in performance

Several Different MT System Outputs

Reference Translation:

hoffman was addicted to drugs, fortunately awaking in a timely manner to begin an acting career

- hoffman was obsessed timely wake up to create a career drug
- hoffman were drug fortunately awakening in a timely manner to create career
- hoffman previously enamored drug. luckily i realized create career
- hoffman was mesmerized by drug but woke up in a timely manner to create career
- hoffmann was obsessed drug, in a timely manner to create a career
- hoffman has fortunately drug come to realize in a timely manner for performing arts to open up the cause

Chinese-English MT06

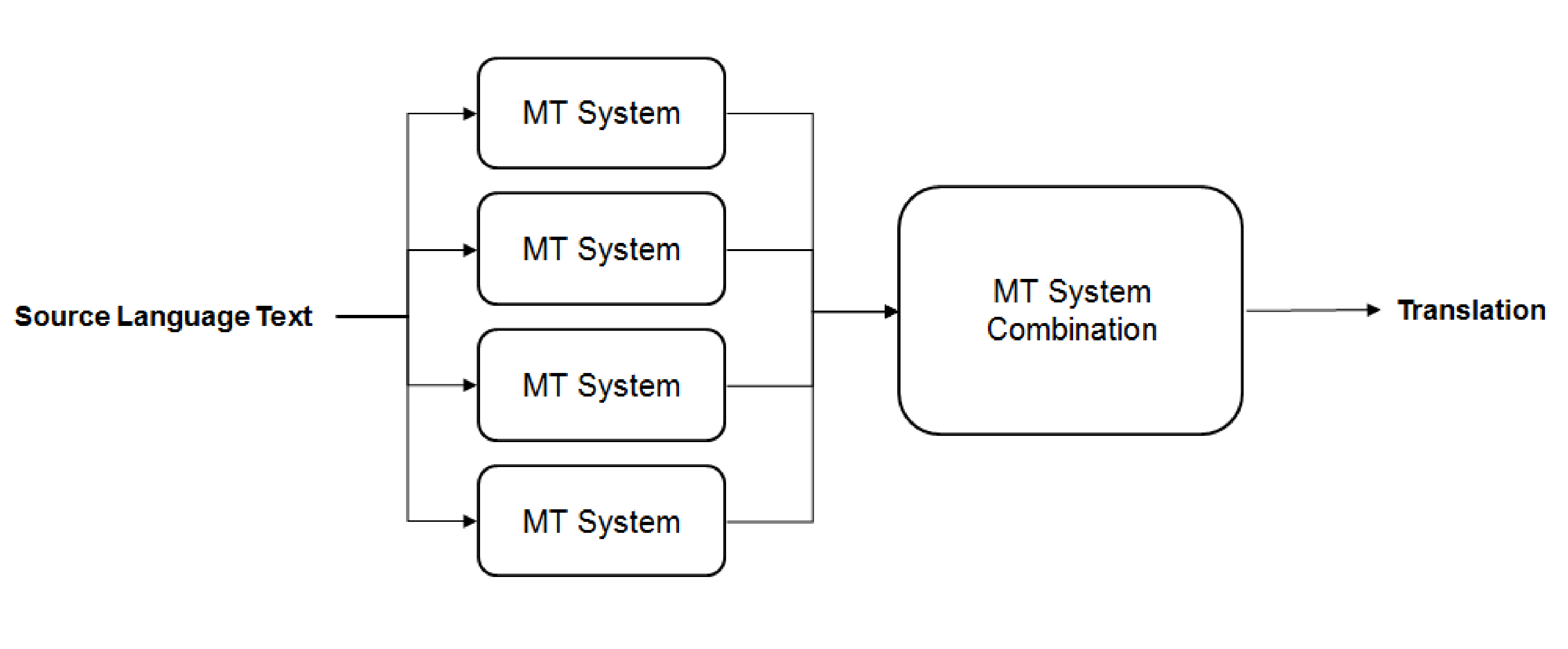
Statistical Phrase Based · Statistical Hierarchical · Example Based

- Translation hypotheses are in order of the systems testset BLEU score

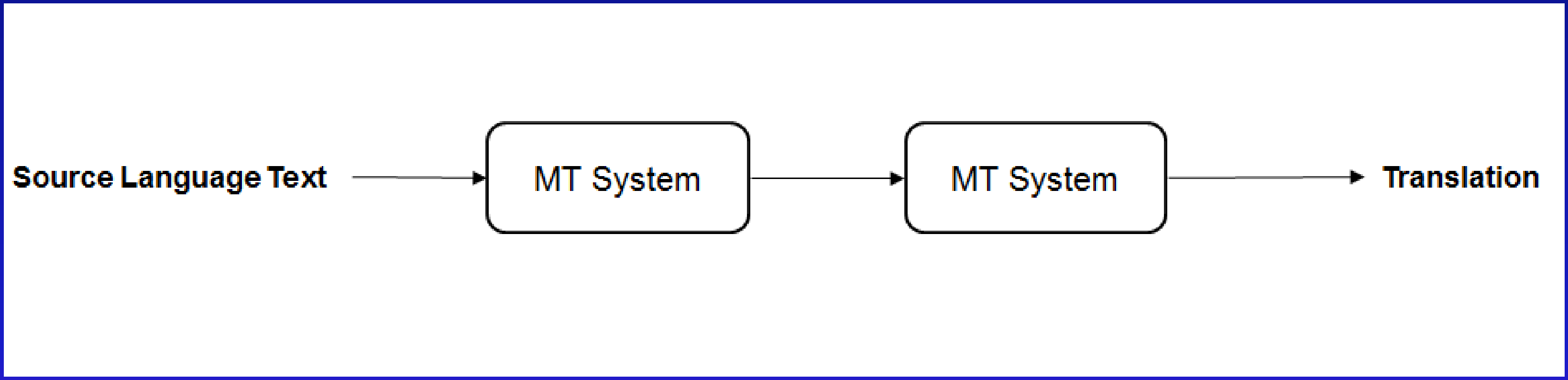
Combination Architecture

- Parallel Combination
  - Run multiple MT systems in parallel, then select or combine their outputs
- Serial Combination
  - Second stage decoding using a different approach
- Model Combination
  - Train separate models, then combine them for joint decoding

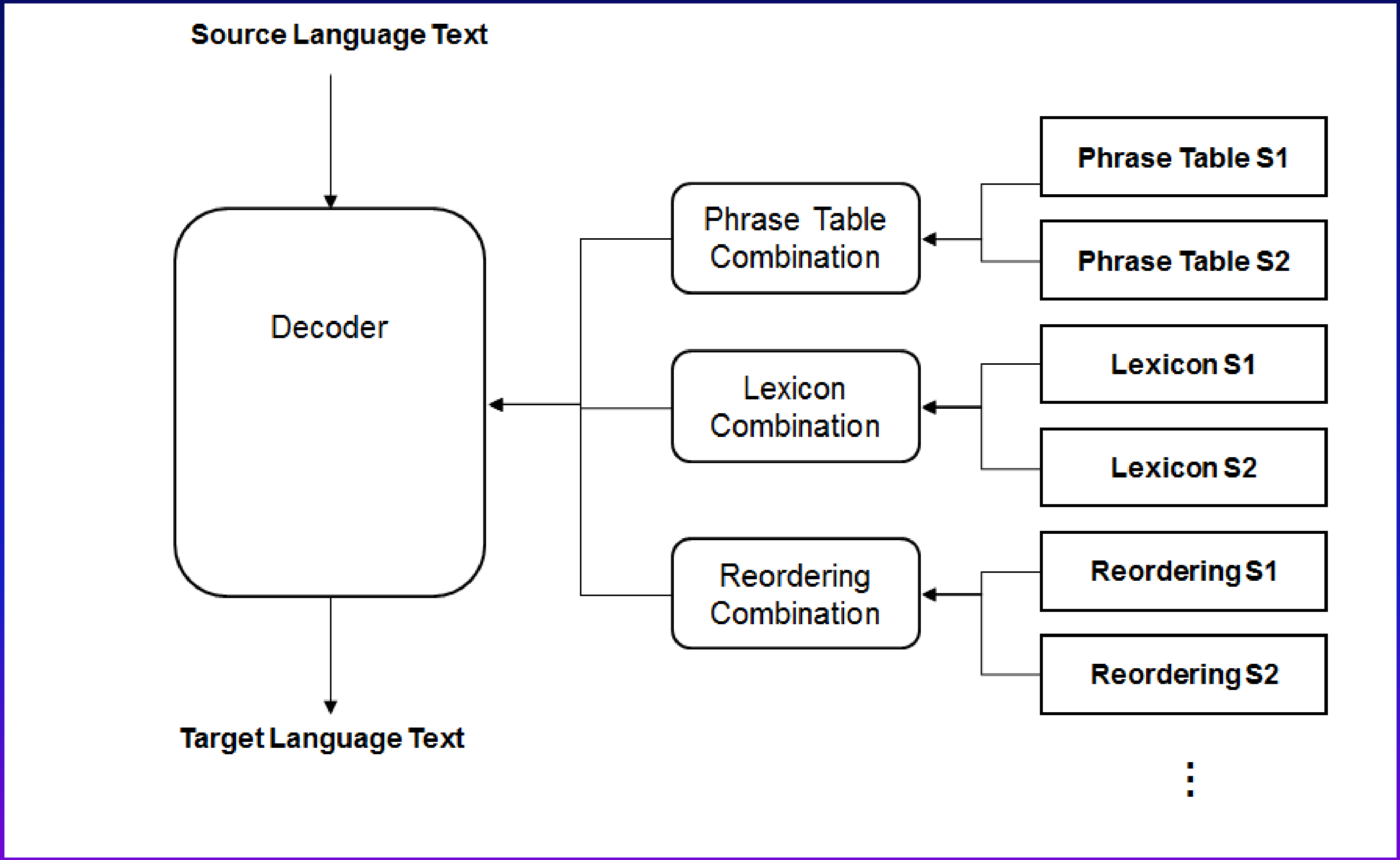
Parallel Combination



Serial Combination



Model Combination



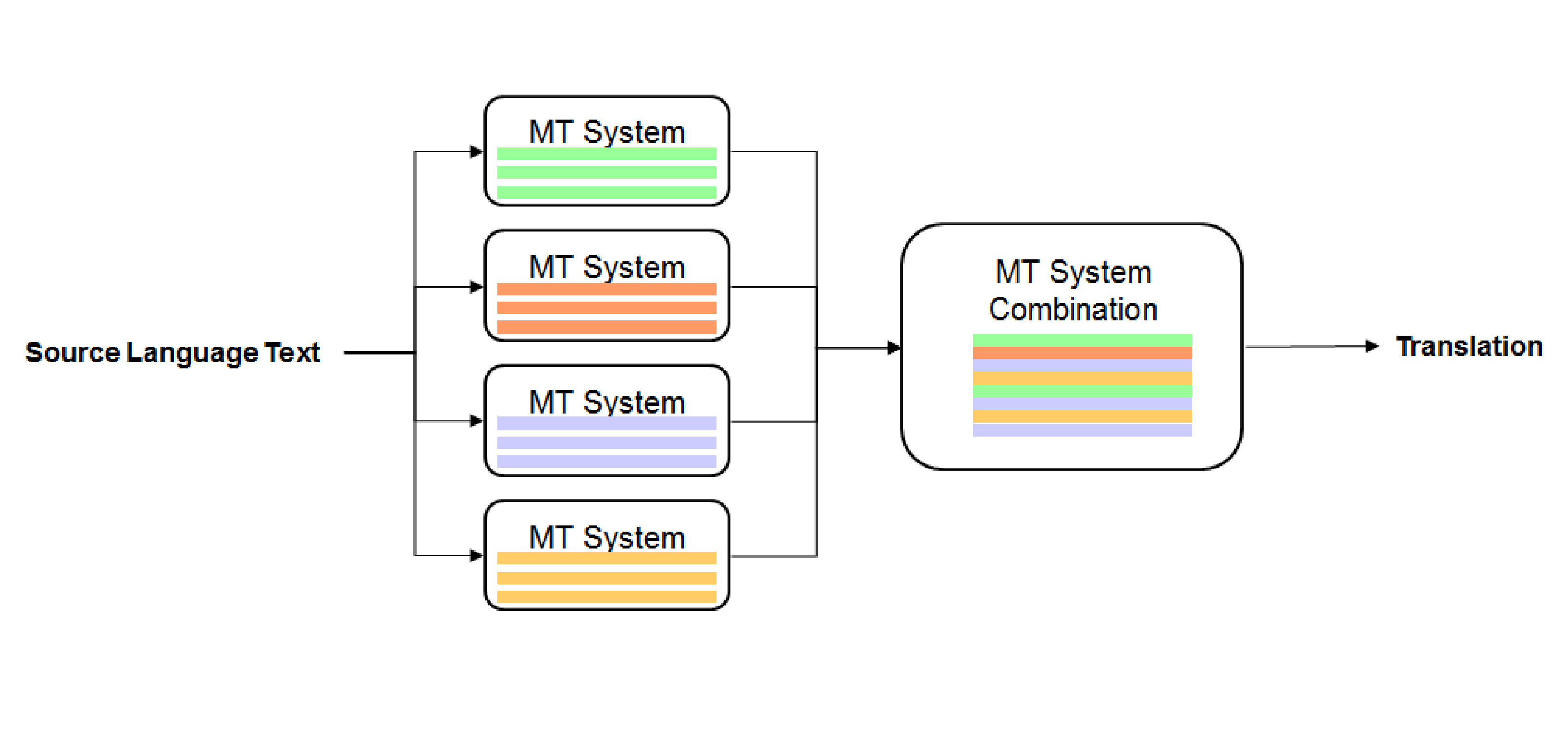
Main Approaches

- Parallel Combination:
  - Hypothesis Selection approaches
  - Lattice Combination
  - Confusion (or Consensus) Networks
  - Alignment-based Synthetic Multi-Engine MT (MEMT)
- Serial Combination:
  - Automated Post-Editing: RBMT + SMT
  - Cross combinations of parallel combinations (GALE)
- Model Combination:
  - Combine lexica, phrase tables, LMs
  - Ensamble decoding (Sarkar et al, 2012)

Hypothesis Selection Approaches

- Main Idea: construct a classifier that given several translations for the same input sentence selects the "best" translation (on a sentence-by-sentence basis)
- Should "beat" a baseline of always picking the system that is best in the aggregate
- Main knowledge sources for scoring the individual translations are standard statistical target-language LMs, confidence scores for each engine, consensus information
- Examples:
  - [Tidhar & Kuessner, 2000]
  - [Hildebrand and Vogel, 2008]

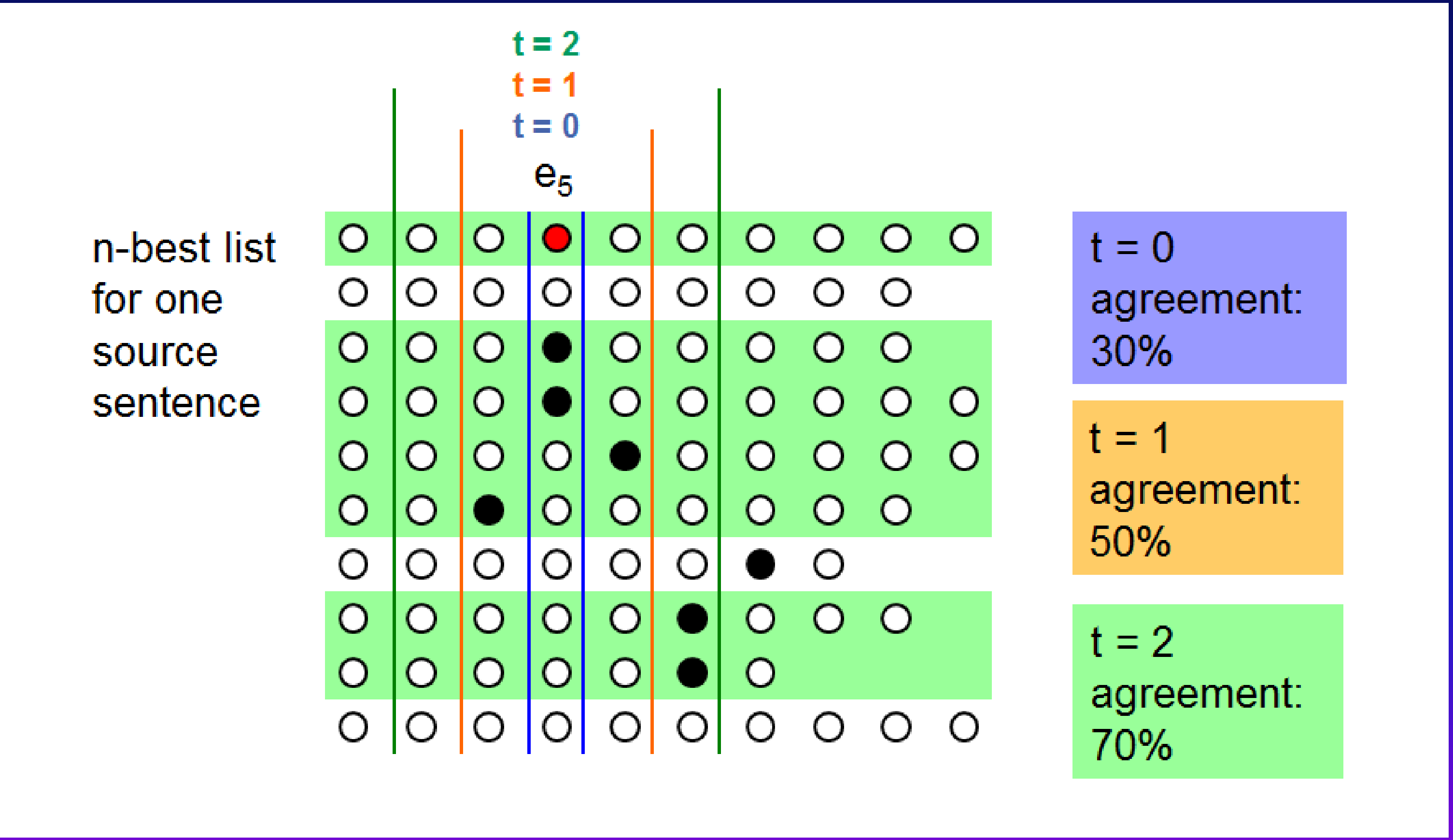
Hypothesis Selection



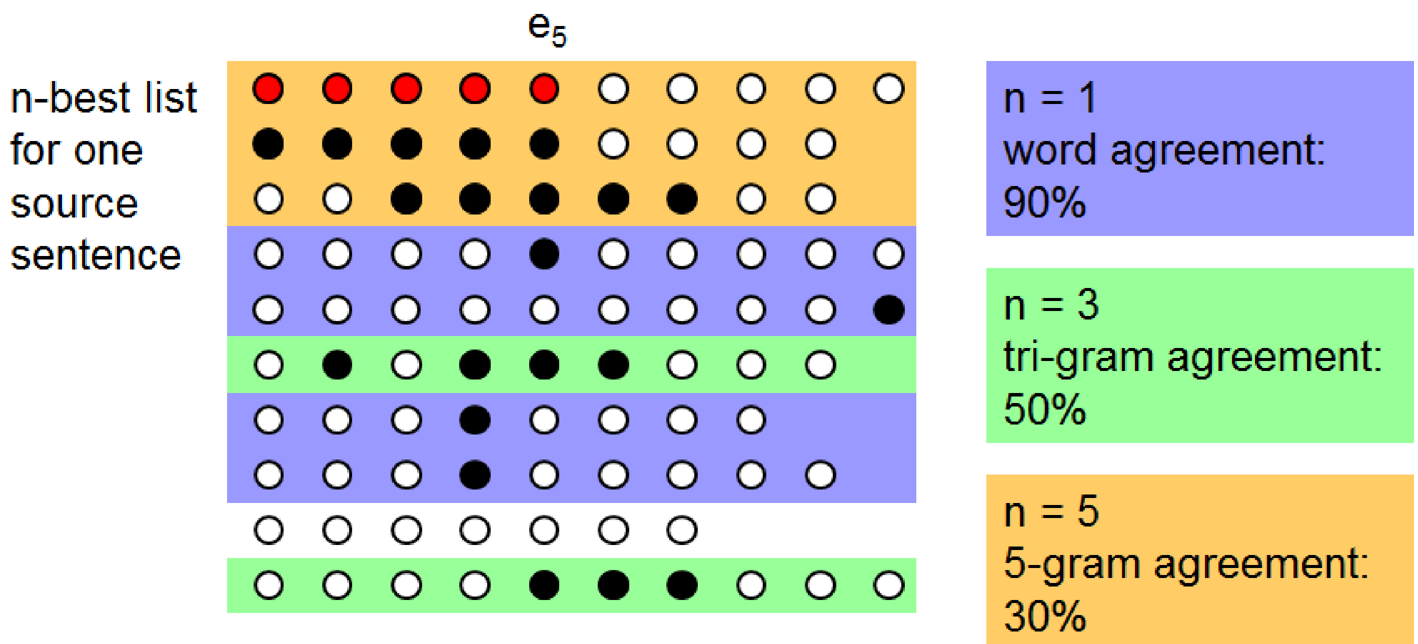
Hypothesis Selection

- Work here at CMU (InterACT) by Silja Hildebrand:
  - Combines n-best lists from multiple MT systems and re-ranks them with a collection of computed features
  - Log-linear feature combination is independently tuned on a development set for max-BLEU
  - Richer set of features than previous approaches, including:
    - Standard n-gram LMs (normalized by length)
    - Lexical Probabilities (from GIZA statistical lexicons)
    - Position-dependent n-best list word agreement
    - Position-independent n-best list n-gram agreement
    - N-best list n-gram probability
    - Aggregate system confidence (based on BLEU)
  - Applied successfully in GALE and WMT-09
  - Improvements of 1-2 BLEU points above the best individual system on average
  - Complimentary to other approaches – is used to select “back-bone” translation for confusion network in GALE

Position-Dependent Word Agreement

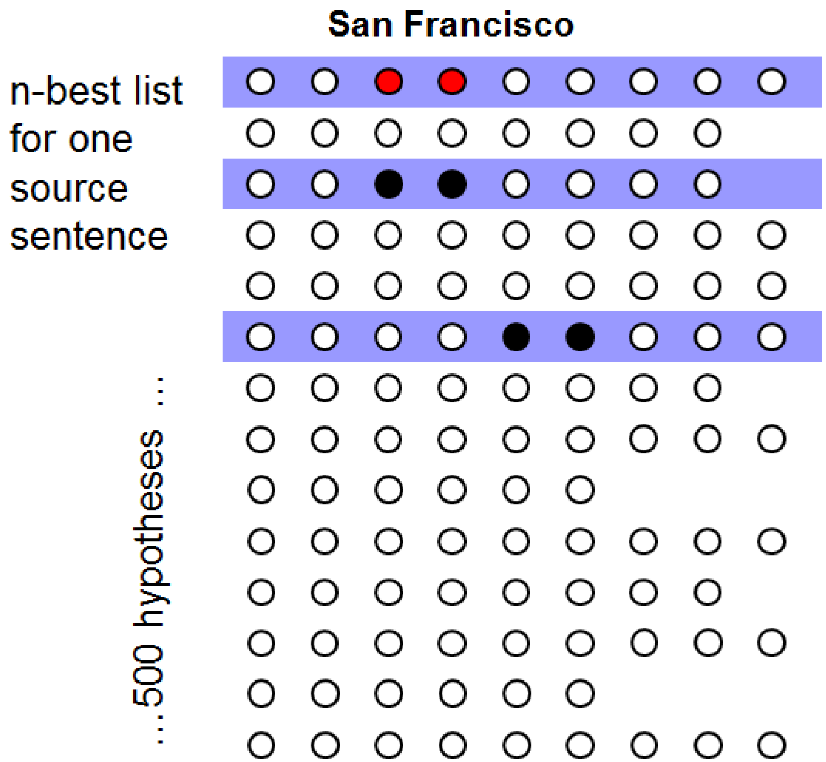


Position-Independent Word Agreement



- Agreement score for n = 1 to 6 as separate features

N-gram Agreement vs. N-gram Probability



n = 2
bi-gram agreement:
0.6%

P(Francisco | San)
= 3 / 3
bi-gram probability:
100%

- LM n-gram probability gives information on word order.

Lattice-based MEMT

- Earliest approach, first tried in CMU's PANGLOSS in 1994, and still active in recent work
- Main Ideas:
  - Multiple MT engines each produce a lattice of scored translation fragments, indexed based on source language input
  - Lattices from all engines are combined into a global comprehensive lattice
  - Joint Decoder finds best translation (or n-best list) from the entries in the lattice

Lattice-based MEMT: Example

El punto de descarge The drop-off point	*se cumplirá en* will comply with	*el puente Agua Fria* The cold Bridgewater
El punto de descarge The discharge point	*se cumplirá en* will self comply in	*el puente Agua Fria* the "Agua Fria" bridge
El punto de descarge Unload of the point	*se cumplirá en* will take place at	*el puente Agua Fria* the cold water of bridge

Lattice-based MEMT

- Main Drawbacks:
  - Requires MT engines to provide lattice output
    → often difficult to obtain!
  - Lattice output from all engines must be compatible: common indexing based on source word positions
    → difficult to standardize!
  - Common TM used for scoring edges may not work well for all engines
  - Decoding does not take into account any reinforcements from multiple engines proposing the same translation for any portion of the input

Consensus Network Approach

- Main Ideas:
  - Collapse the collection of linear strings of multiple translations into a minimal consensus network ("sausage" graph) that represents a finite-state automaton
  - Edges that are supported by multiple engines receive a score that is the sum of their contributing confidence scores
  - Decode: find the path through the consensus network that has optimal score
  - Examples:
    - [Bangalore et al, 2001]
    - [Rosti et al, 2007]

Consensus Network Example

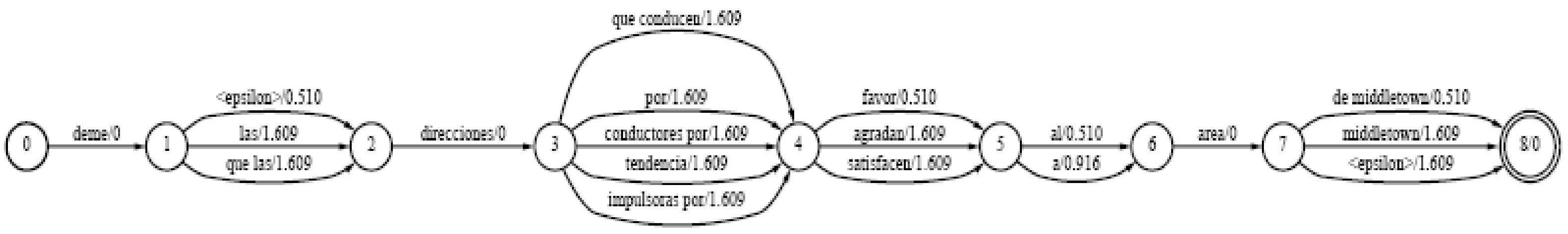
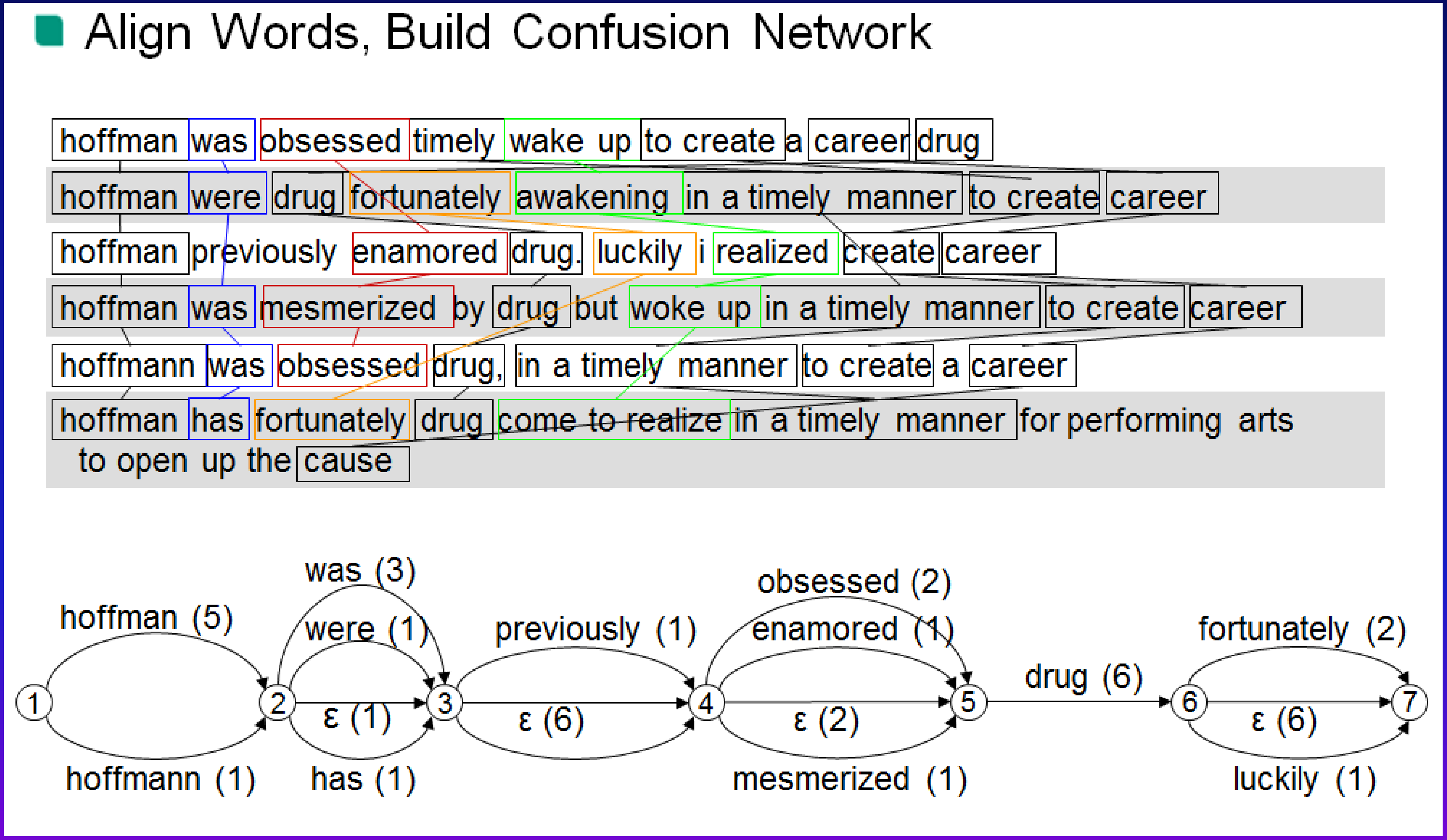


Fig. 4. Lattice representation of the result of the multiple alignment. The weights on the arcs are negative logarithm of the probability that word.

Confusion Network Approaches

- Similar in principle to the Consensus Network approach
  – Collapse the collection of linear strings of multiple translations into minimal confusion network(s)
- Main Ideas and Issues:
  – Aligning the words across the various translations:
    - Can be aligned using TER, ITGs, statistical word alignment
  – Word Ordering: picking a “back-bone” translation
    - One backbone? Try each original translation as a backbone?
  – Decoding Features:
    - Standard n-gram LMs, system confidence scores, agreement
  – Decode: find the path through the consensus network that has optimal score
- Developed and used extensively in GALE (also WMT)
- Nice gains in translation quality: 1-4 BLEU points

Confusion Network Construction



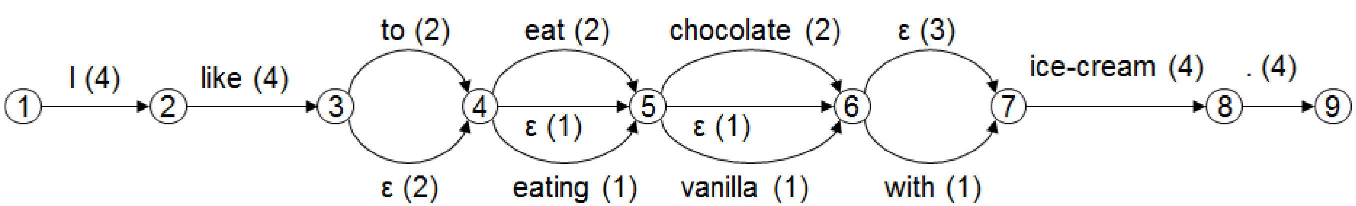
Confusion Network Decoding

I like eating chocolate ice-cream.
I like to eat vanilla ice-cream.
I like to eat ice-cream with chocolate.
I like ice-cream.

choose as skeleton

I	like	ε	eating	chocolate	ε	ice-cream	.
I	like	to	eat	vanilla	ε	ice-cream	.

skeleton determines word order



I like to eat chocolate ice-cream.

Confusion Networks - Challenges

- Word alignment
  - TER alignment (Translation Edit Rate)
  - ITG based alignment (Inversion Transduction Grammar) - invWER
  - Use morphology, synonyms, POS tag
  - Go to phrases
    - Difficult without source-target phrase alignment available
- Double translations
- Dropped words
- Pairwise vs. incremental alignment
  - Next hypothesis is aligned to the existing network, not to the skeleton
  - Order of adding hypothesis does make a difference, e.g. use increasing TER/decreasing BLEU of the system

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 ten years under research funding from several government grants (DARPA, DoD and NSF)

Alignment-based MEMT

Two Stage Approach:

1. Identify common 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. Identify common 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

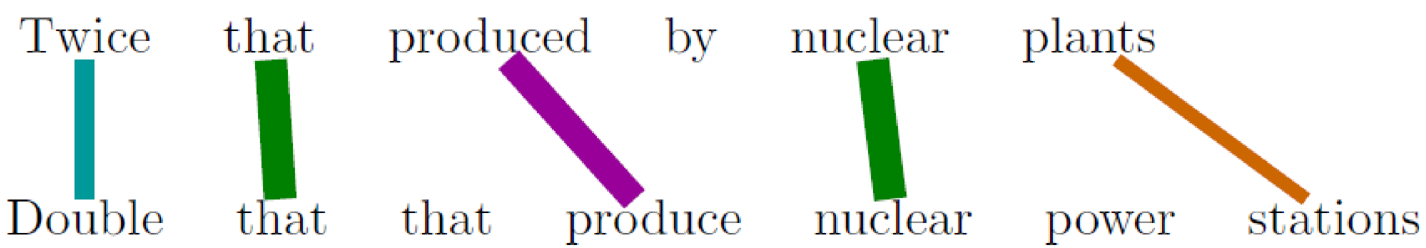
MEMT: the afghan authorities announced on Saturday the formation of four intergovernmental committees

The String Alignment Matcher

- Developed as a component in the METEOR Automatic MT Evaluation metric
- Finds maximal alignment match with minimal “crossing branches”
- Allows alignment of:
  - Identical words
  - Morphological variants of words
  - Synonymous words (based on WordNet synsets)
  - Paraphrases
- Implementation: approximate single-pass search algorithm for best match using pruning of sub-optimal sub-solutions

MEMT Alignment

Match surface, stems, WordNet synsets, and automatic paraphrases
Minimize crossing alignments

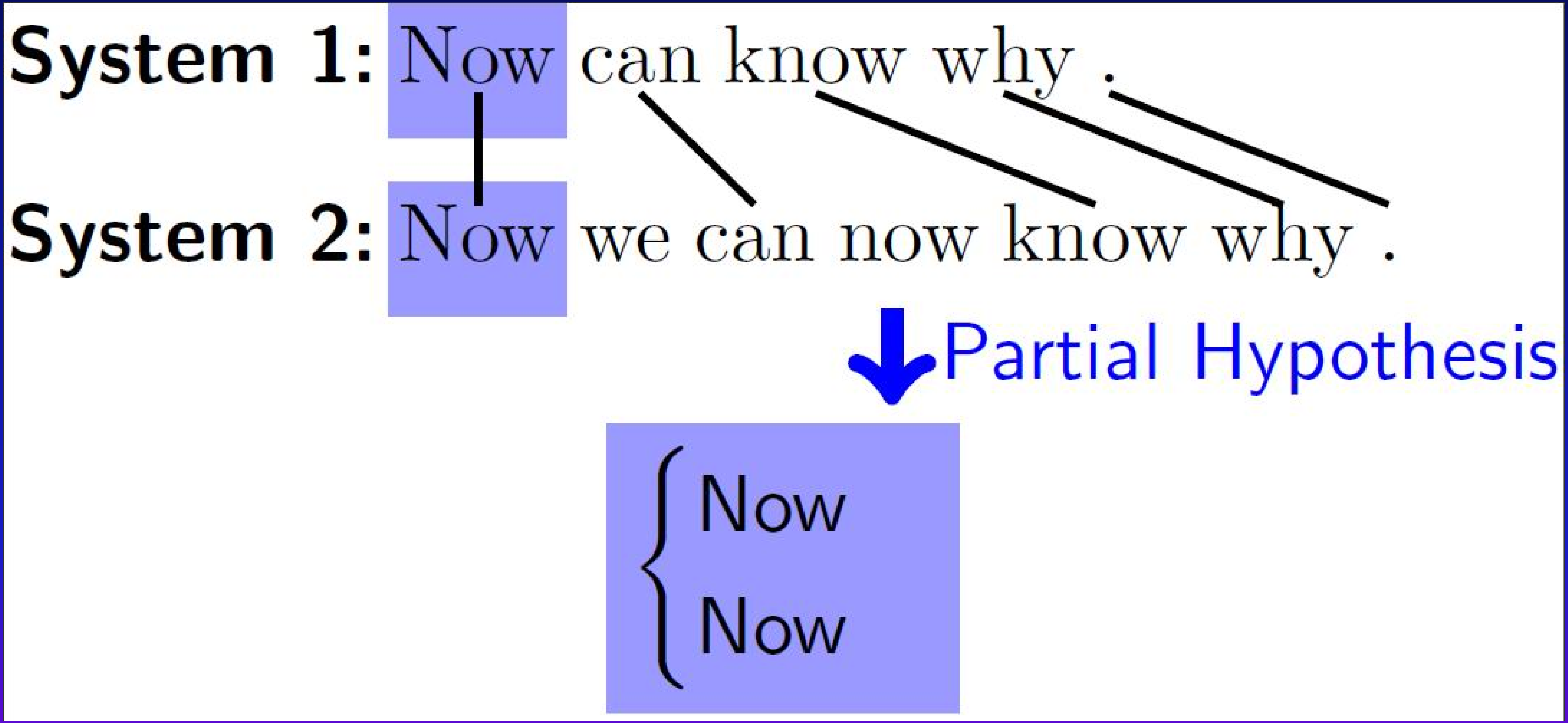


Lavie and Agarwal, METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments, WMT 2007.

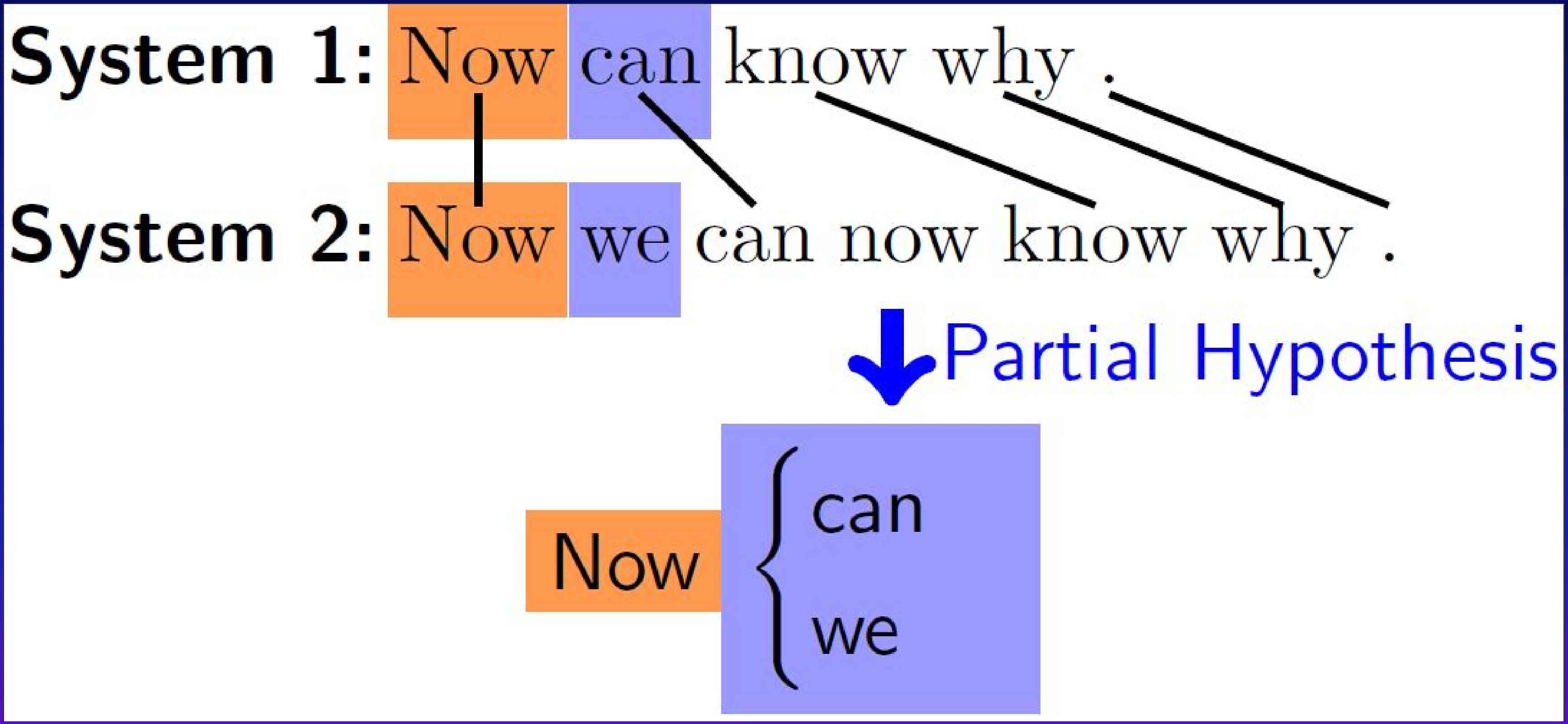
The MEMT Decoder Algorithm

- 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
- Sentences are assumed mostly synchronous:
  - Each word is either *aligned* with another word or is an *alternative* of another word
- Extending a partial hypothesis with a word "pulls" and "uses" its aligned words with it, and marks its alternatives as "used"
- Partial hypotheses are scored and ranked
- Pruning and re-combination
- Hypothesis can end if any original system proposes an end of sentence as next word

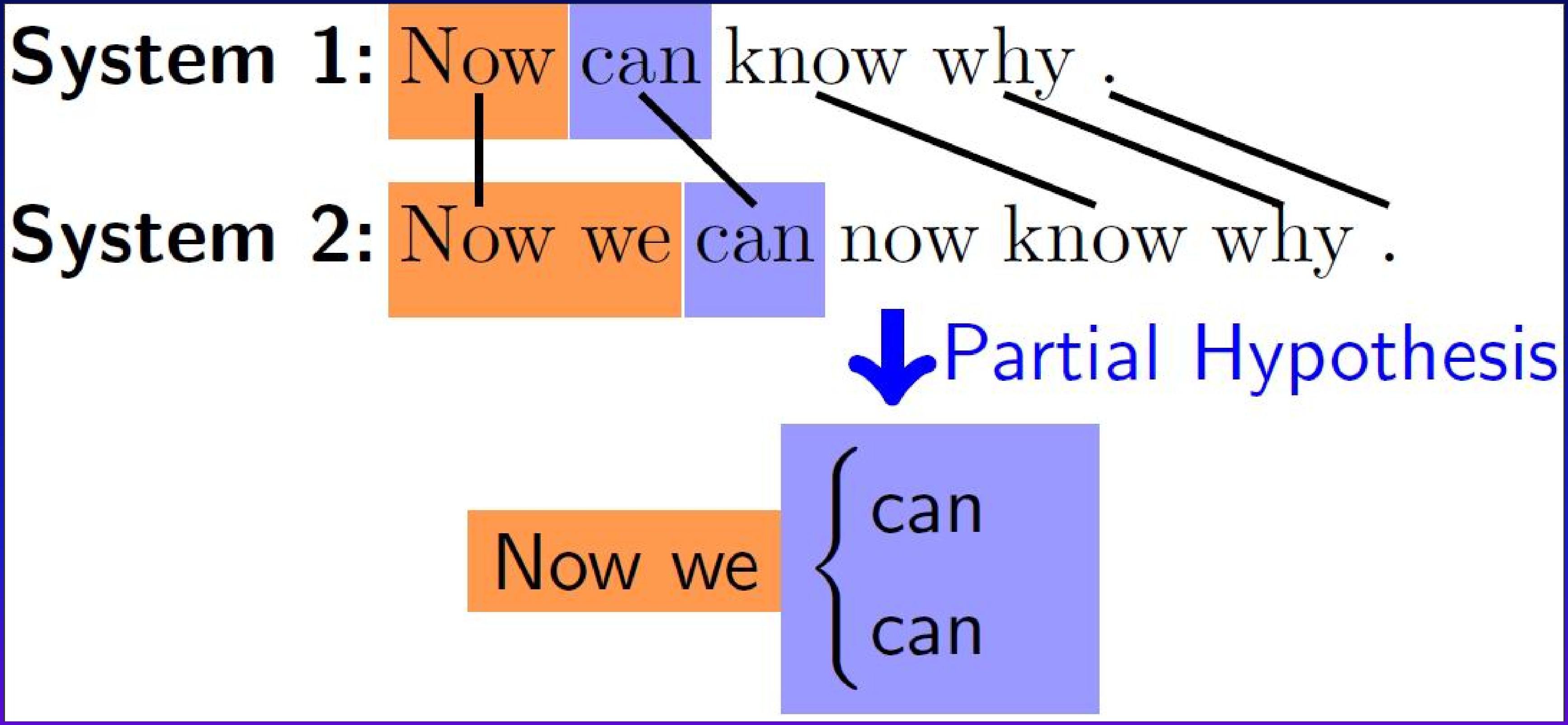
Decoding Example



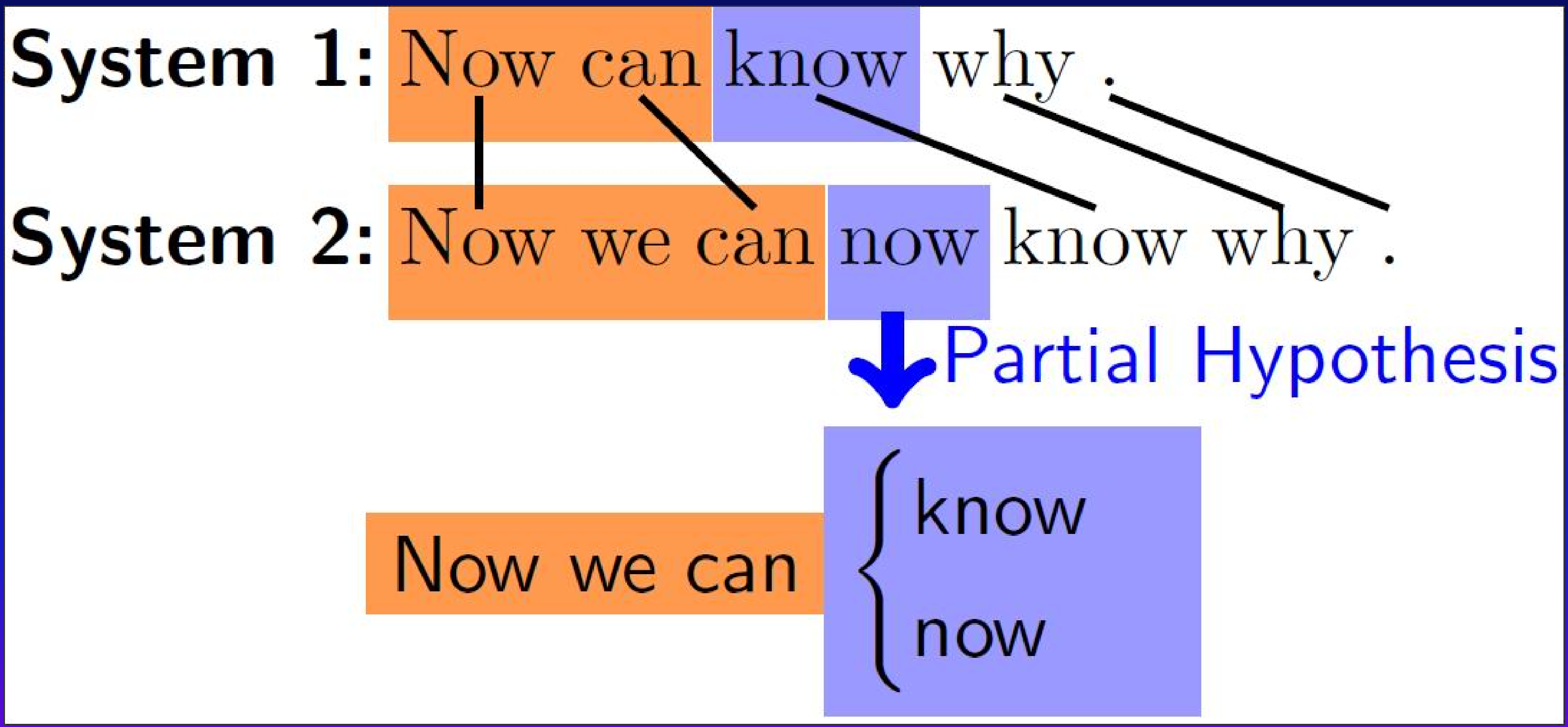
Decoding Example



Decoding Example



Decoding Example



Scoring MEMT Hypotheses

- Features:
  - N-gram Language Model score based on filtered large-scale target language LM
  - OOV feature
  - 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

System 1: Supported Proposal of France

System 2: Support for the Proposal of France

↓ Hypothesis

Hypothesis: Support for Proposal of France

↓ Count

	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

Performance Results on 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.

Performance Results on 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

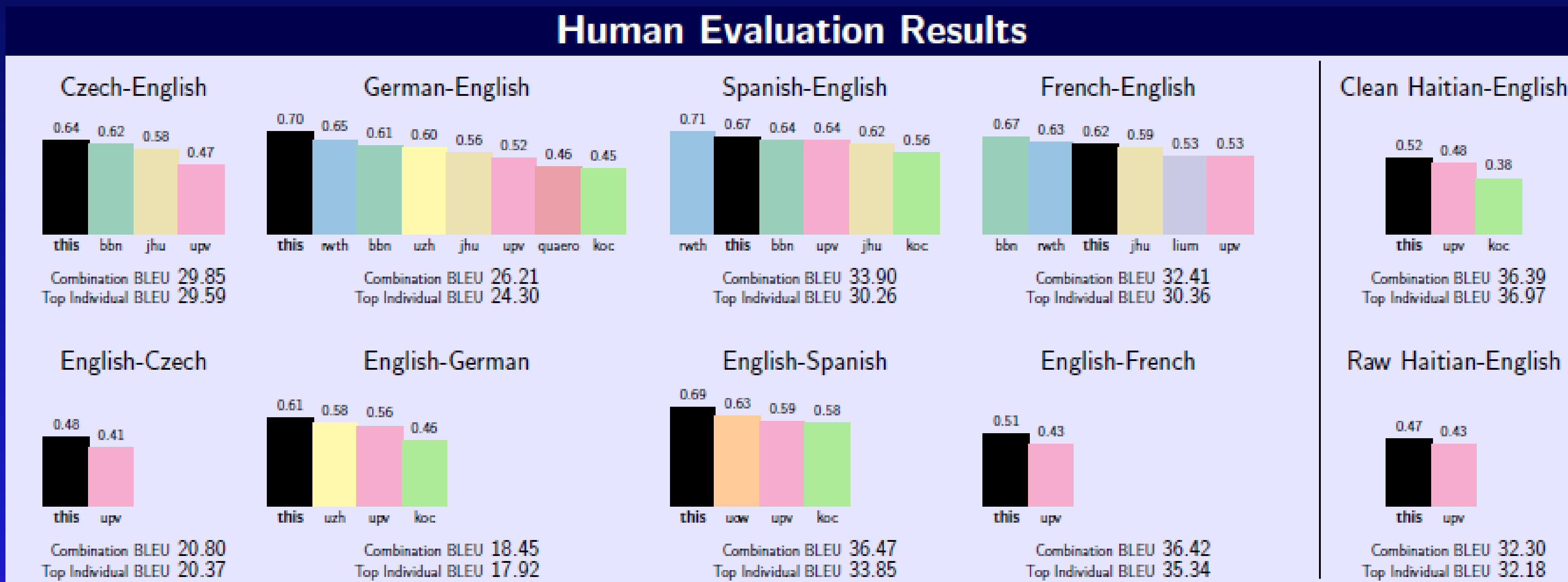
System	≥others
RWTH-COMBO ●	0.75
CMU-HEA-COMBO ●	0.74
UEDIN	0.70
KOC-COMBO ●	0.68
UPV-COMBO	0.66
RALI ⋆	0.66
LIMSI	0.66
RWTH	0.63
CAMBRIDGE	0.63

Performance Results on 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

Performance Results on 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

References

- 1994, Frederking, R. and S. Nirenburg. “Three Heads are Better than One”. In Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP-94), Stuttgart, Germany.
- 2000, Tidhar, Dan and U. Kessner. “Learning to Select a Good Translation”. In Proceedings of the 17th International Conference on Computational Linguistics (COLING-2000), Saarbrcken, Germany.
- 2001, Bangalore, S., G. Bordel, and G. Riccardi. “Computing Consensus Translation from Multiple Machine Translation Systems”. In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, Italy.
- 2005, Jayaraman, S. and A. Lavie. "Multi-Engine Machine Translation Guided by Explicit Word Matching" . In Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT-2005), Budapest, Hungary, May 2005.
- 2007, Rosti, A-V. I., N. F. Ayan, B. Xiang, S. Matsoukas, R. Schwartz and B. J. Dorr. “Combining Outputs from Multiple Machine Translation Systems”. In Proceedings of *NAACL-HLT-2007 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, April 2007, Rochester, NY; pp.228-235
- 2008, Hildebrand, A. S. and S. Vogel. “Combination of Machine Translation Systems via Hypothesis Selection from Combined N-best Lists”. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA-2008)*, Waikiki, Hawai’i, October 2008; pp.254-261
- 2009, Heafield, K., G. Hanneman and A. Lavie. "Machine Translation System Combination with Flexible Word Ordering" . In Proceedings of the Fourth Workshop on Statistical Machine Translation at the 2009 Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009), Athens, Greece, March 2009.
- 2010, Heafield, K. and A. Lavie. "Voting on N-grams for Machine Translation System Combination" . In Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA-2010), Denver, Colorado, November 2010.

Questions?