Modelling Sentence Generation from Sum of Word Embedding Vectors as a Mixed Integer Programming Problem

Moving from a 300 dimentional meaning space, back to human sentences

Lyndon White,
Roberto Togneri, Wei Liu, Mohammed Bennamoun

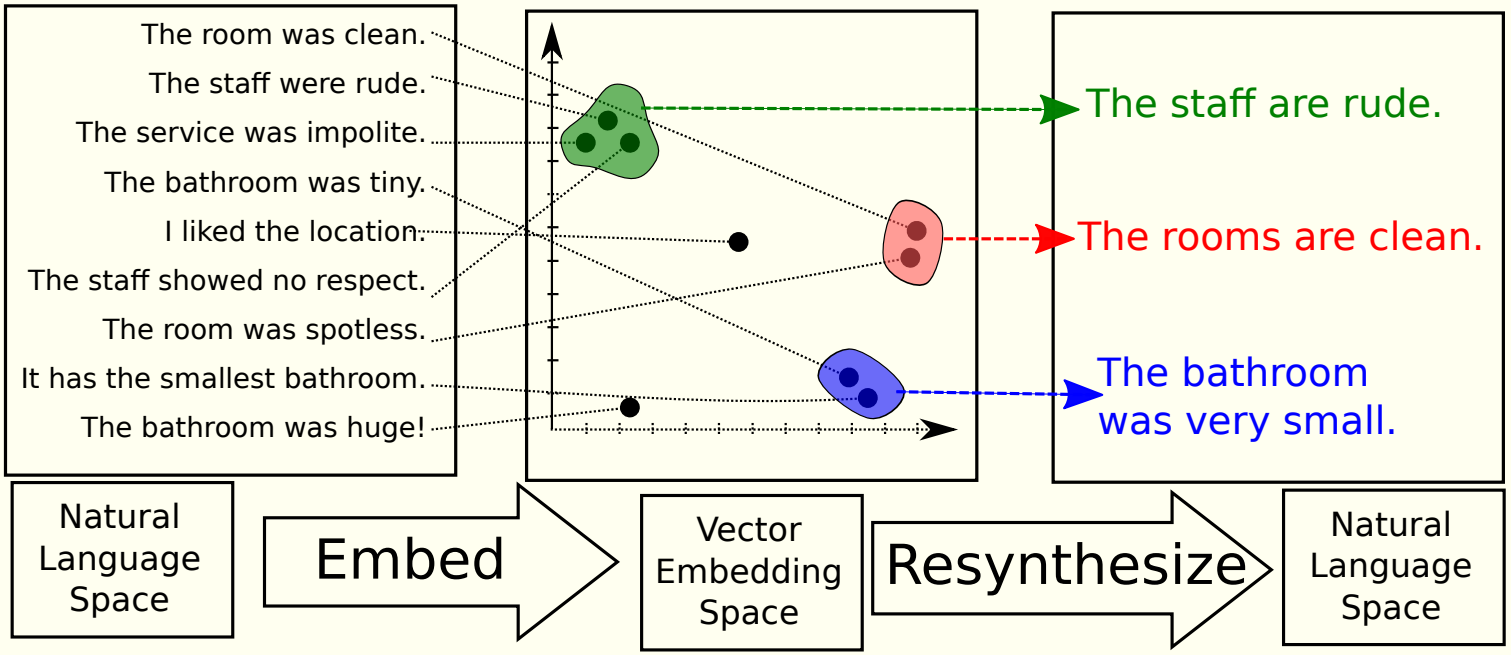
School of EE&C Engineering
The University of Western Australia

We have turned sentences into numeric vectors, now we want to turn them back.

1. It was the best of times, it was the worst of times
2. $[0.79, 1.27, 0.28, \ldots, 1.29]$
3. It was the worse of times, it was the best of times

The use of an ideal generation system is in allowing manipulation in the vector domain

Input Sentences Manipulate Numbers Output Sentences



Related problems

Related problems include

- Sequence Memorisation
- Sequence-Sequence Learning

- Sentence generation is different from either.
- Going from a vector representation that only encodes meaning
- Rather than from one that encodes memory of meaning.

We have turned sentences into numeric vectors, now we want to turn them back.

1. It was the best of times, it was the worst of times
2. $[0.79, 1.27, 0.28, \ldots, 1.29]$
3. It was the worse of times, it was the best of times

Related work: Dependancy Tree RAE

Method

- Input as tree structured of word embeddings combined with single layer nets
- Output same structure
- Use Back Propagating Through Structure in training

Mohit Iyyer, Jordan Boyd-Graber, and Hal Daumé III. "Generating Sentences from Semantic Vector Space Representations". In: *NIPS Workshop on Learning Semantics*. 2014.

Related work: Dependancy Tree RAE

Method

- Input as tree structured of word embeddings combined with single layer nets
- Output same structure
- Use Back Propagating Through Structure in training

- Produces fairly clean paraphrases
- Requires output tree structure to be provided

Mohit Iyyer, Jordan Boyd-Graber, and Hal Daumé III. "Generating Sentences from Semantic Vector Space Representations". In: *NIPS Workshop on Learning Semantics*. 2014.

Related work: LSTM + Variational Autoencoder

Method

- LSTM encoding step
- Variational Autoencoder Representation step
- LSTM decoding step

Samuel R Bowman et al. “Generating Sentences from a Continuous Space”. In: *International Conference on Learning Representations (ICLR) Workshop* (2016).

Related work: LSTM + Variational Autoencoder

Method

- LSTM encoding step
- Variational Autoencoder Representation step
- LSTM decoding step

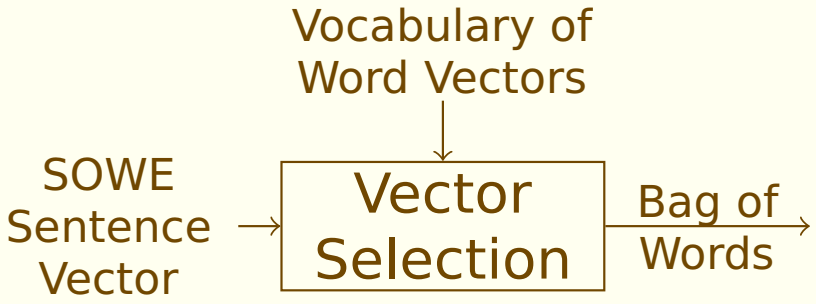
- Smooth “deformation” between sentences
- Several other uses beyond just generation
- No demonstration on sentences with more than 8 words

Samuel R Bowman et al. “Generating Sentences from a Continuous Space”. In: *International Conference on Learning Representations (ICLR) Workshop* (2016).

We have turned sentences into numeric vectors, now we want to turn them back.

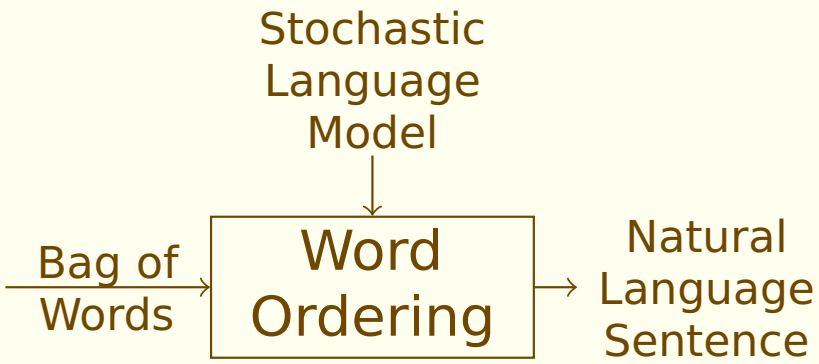
1. It was the best of times, it was the worst of times
2. $[0.79, 1.27, 0.28, \ldots, 1.29]$
3. It was the worse of times, it was the best of times

We broke the problem down into two subproblems.



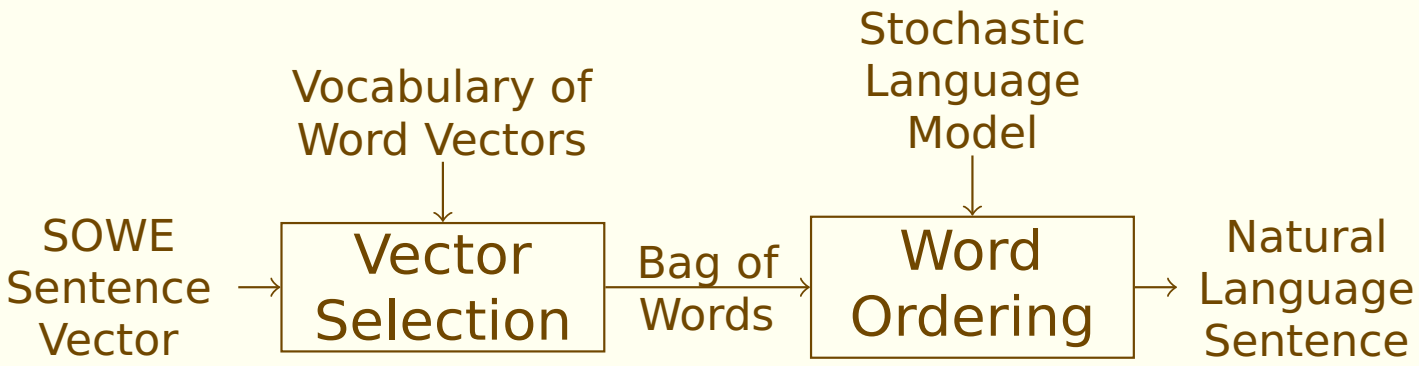
Vector Selection: Select which word vectors go into the sum

We broke the problem down into two subproblems.



Word Ordering: Find them most likely order of words

We broke the problem down into two subproblems.



They are however both NP-Hard

We get a bag of words by minimising an objective function

Find the inclusion vector $\tilde{c} = [c_1, c_2, \ldots c_n] \in \mathbb{N}_0^n$ to minimise $d(\tilde{s}, \sum_{\tilde{x}_j \in \mathcal{V}} c_j \tilde{x}_j)$

We get a bag of words by minimising an objective function

Find the inclusion vector $\tilde{c} = [c_1, c_2, \ldots c_n] \in \mathbb{N}_0^n$ to minimise $d(\tilde{s}, \sum_{\tilde{x}_j \in \mathcal{V}} c_j \tilde{x}_j)$

Input Vector $\tilde{s} = [0.79, 1.27, 0.28, \ldots, 1.29]$

We get a bag of words by minimising an objective function

Find the inclusion vector $\tilde{c} = [c_1, c_2, \ldots c_n] \in \mathbb{N}_0^n$ to minimise $d(\tilde{s}, \sum_{\tilde{x}_j \in \mathcal{V}} c_j \tilde{x}_j)$

Input Vector $\tilde{s} = [0.79, 1.27, 0.28, \ldots, 1.29]$

Vector Selection

$$\begin{aligned}\sum_{\tilde{x}_j \in \mathcal{V}} c_j \tilde{x}_j) = \quad & 1 \times [0.19, 0.50, 0.14, \ldots, 0.59] \\ & + 2 \times [-0.15, 0.19, 0.03, \ldots, -0.17] \\ & + \quad \ldots \\ & + 0 \times [0.19, 2.10, 1.34, \ldots, 1.20] \\ & + 1 \times [0.79, 1.27, 0.28, \ldots, 1.29]\end{aligned}$$

We get a bag of words by minimising an objective function

Find the inclusion vector $\tilde{c} = [c_1, c_2, \ldots c_n] \in \mathbb{N}_0^n$ to minimise $d(\tilde{s}, \sum_{\tilde{x}_j \in \mathcal{V}} c_j \tilde{x}_j)$

Input Vector $\tilde{s} = [0.79, 1.27, 0.28, \ldots, 1.29]$

Vector Selection

$$\begin{aligned}\sum_{\tilde{x}_j \in \mathcal{V}} c_j \tilde{x}_j) = \quad & 1 \times [0.19, 0.50, 0.14, \ldots, 0.59] \\ & + 2 \times [-0.15, 0.19, 0.03, \ldots, -0.17] \\ & + \quad \ldots \\ & + 0 \times [0.19, 2.10, 1.34, \ldots, 1.20] \\ & + 1 \times [0.79, 1.27, 0.28, \ldots, 1.29]\end{aligned}$$

BOW {best: 1,times: 2, worst: 1, it: 2, of: 2, the: 2, was: 2„ : 1}

How to solve objective function? Greedily

Find the inclusion vector $\tilde{c} = [c_1, c_2, \ldots c_n] \in \mathbb{N}_0^n$ to minimise $d(\tilde{s}, \sum_{\tilde{x}_j \in \mathcal{V}} c_j \tilde{x}_j)$

- Similarities to Knapsack family of problems.

Lyndon White et al. “Generating Bags of Words from the Sums of their Word Embeddings”. In: *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. 2016.

How to solve objective function? Greedily

Find the inclusion vector $\tilde{c} = [c_1, c_2, \ldots c_n] \in \mathbb{N}_0^n$ to minimise $d(\tilde{s}, \sum_{\tilde{x}_j \in \mathcal{V}} c_j \tilde{x}_j)$

- Similarities to Knapsack family of problems.
- Very high dimensionality of selection vector
  - n is given by vocabulary size ($n = |\mathcal{V}|$)
  - $\approx 170,000$ for Books Corpus

Lyndon White et al. "Generating Bags of Words from the Sums of their Word Embeddings". In: *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. 2016.

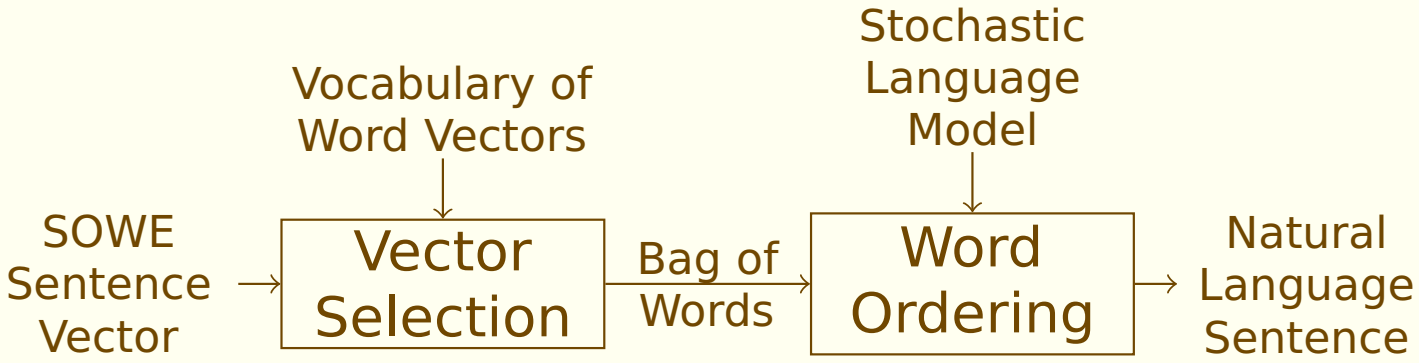
How to solve objective function? Greedily

Find the inclusion vector $\tilde{c} = [c_1, c_2, \ldots c_n] \in \mathbb{N}_0^n$ to minimise $d(\tilde{s}, \sum_{\tilde{x}_j \in \mathcal{V}} c_j \tilde{x}_j)$

- Similarities to Knapsack family of problems.
- Very high dimensionality of selection vector
  - n is given by vocabulary size ($n = |\mathcal{V}|$)
  - $\approx 170,000$ for Books Corpus
- A Greedy Algorithm is linear time in n

Lyndon White et al. "Generating Bags of Words from the Sums of their Word Embeddings". In: *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. 2016.

We broke the problem down into two subproblems.



Vector Selection: Select which word vectors go into the sum

Word Ordering: Find them most likely order of words

Now that we have a bag of words, we need to order them to get a sentence.

Find the most-likely ordering, of the bag of words.

Input Vector $[0.79, 1.27, 0.28, \ldots, 1.29]$

Bag of Words `{best: 1,times: 2, worst: 1, it: 2, of: 2, the: 2, was: 2„ : 1}`

Output Sentence It was the worse of times, it was the best of times

A language model tells us the probability of a word sequence.

- The language model is based on corpus statistics.
- We use a trigram language model:
  $P(W_3 = buns \mid W_1 = hot, W_2 = crossed)$

A language model tells us the probability of a word sequence.

- The language model is based on corpus statistics.
- We use a trigram language model:
  $P(W_3 = buns \mid W_1 = hot, W_2 = crossed)$
- By making a Markov assumption, we can use the trigram probabilities to estimate the probability of any word sequence.

A language model tells us the probability of a word sequence.

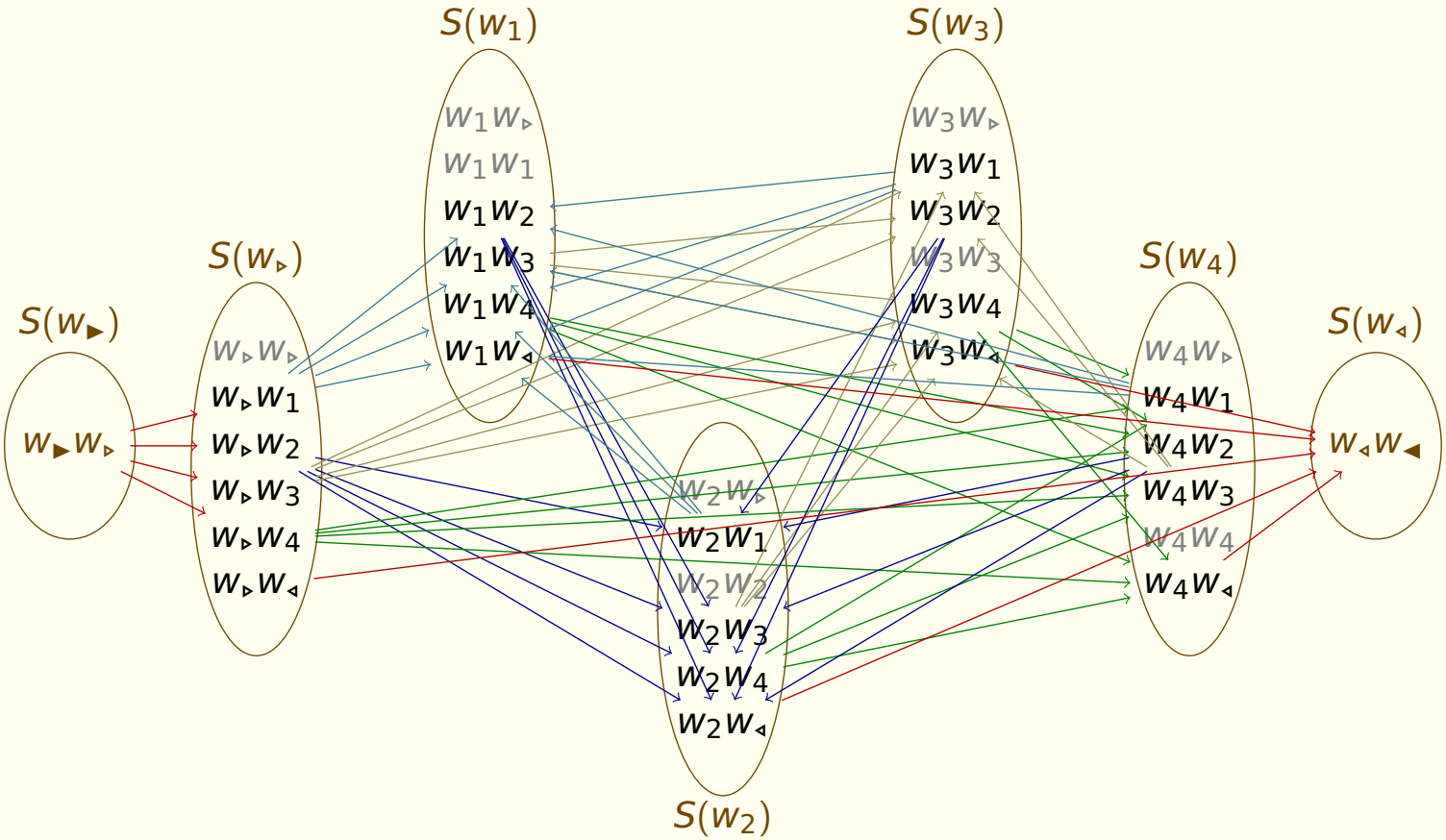
- The language model is based on corpus statistics.
- We use a trigram language model: $P(W_3 = buns \mid W_1 = hot, W_2 = crossed)$
- By making a Markov assumption, we can use the trigram probabilities to estimate the probability of any word sequence.
- Bayesian Chain Rule: $P([w_1, w_2, w_3, w_4, w_5]) =$

$$P(w_1, w_2) \cdot P(w_3|w_1, w_2) \cdot P(w_4|w_2, w_3) \cdot P(w_5|w_3, w_2)$$

We can formulate word ordering as a Mixed Integer Programming (MIP) problem

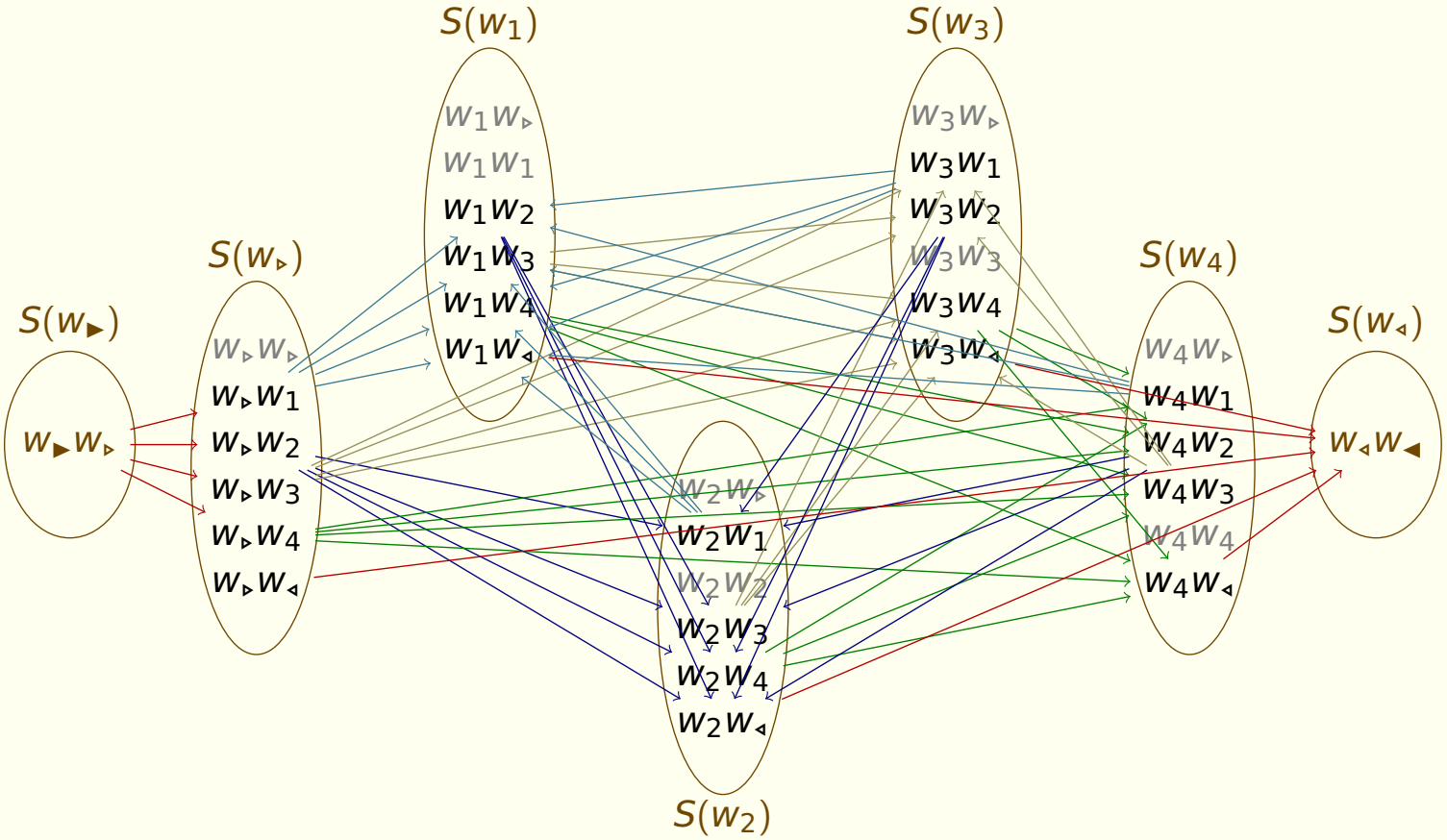
- There exist very fast MIP solvers.
- This gave multiple orders of magnitude improvement over best first search.
- and even over incomplete beam search.

Word Sequencing as a Travelling Salesman Problem



Matic Horvat and William Byrne. "A Graph-Based Approach to String Regeneration." In: *EACL*. 2014, pp. 85–95.

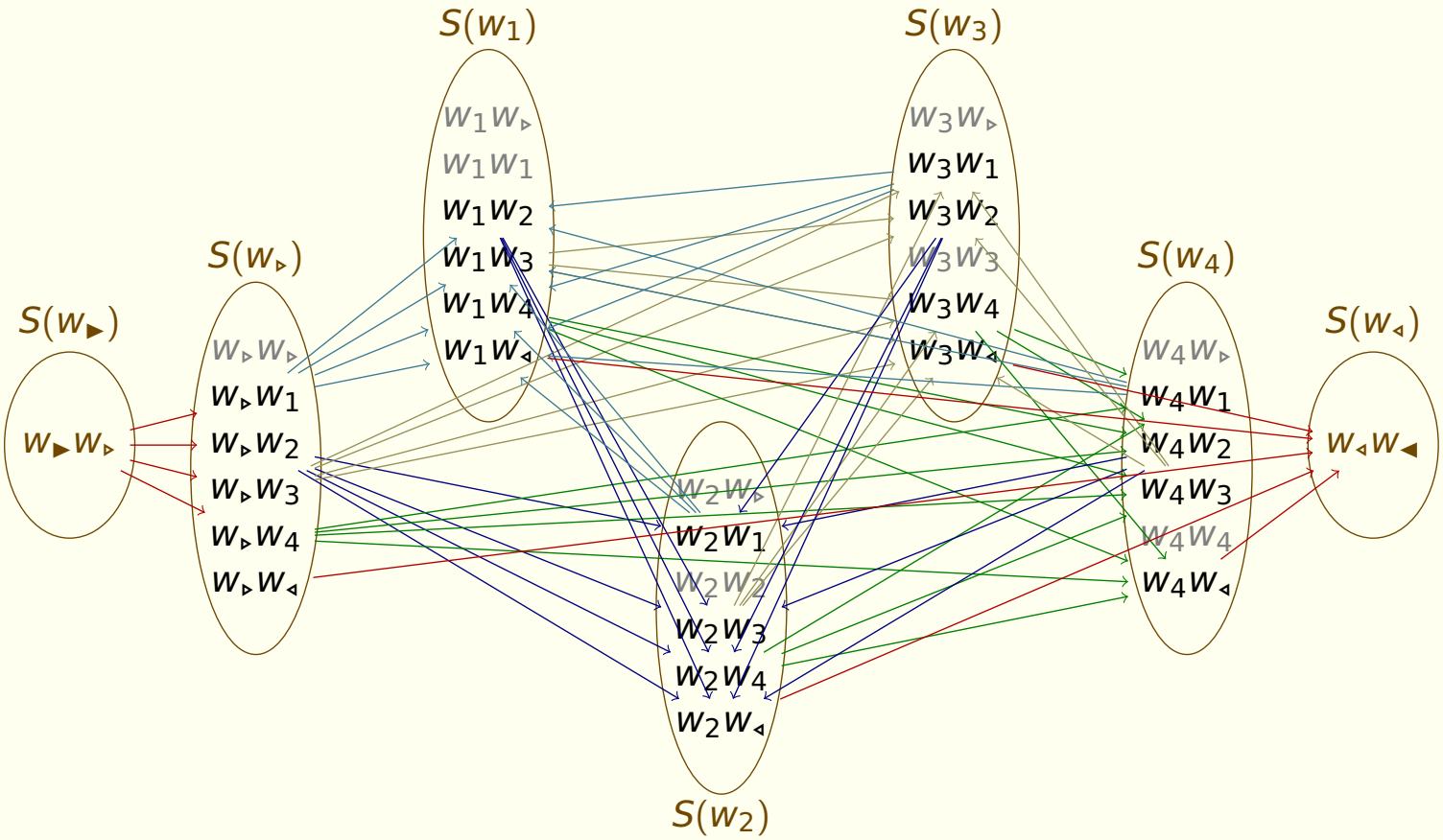
Edge Constraints



Markov Consistency: $(w_a w_b) \rightarrow (w_c w_d) \iff w_b == w_c$

$$\tau[\langle w_i, w_j \rangle, \langle w_j, w_k \rangle] = \begin{cases} 1 & \text{if transition from} \\ & \langle w_i, w_j \rangle \rightarrow \langle w_j, w_k \rangle \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

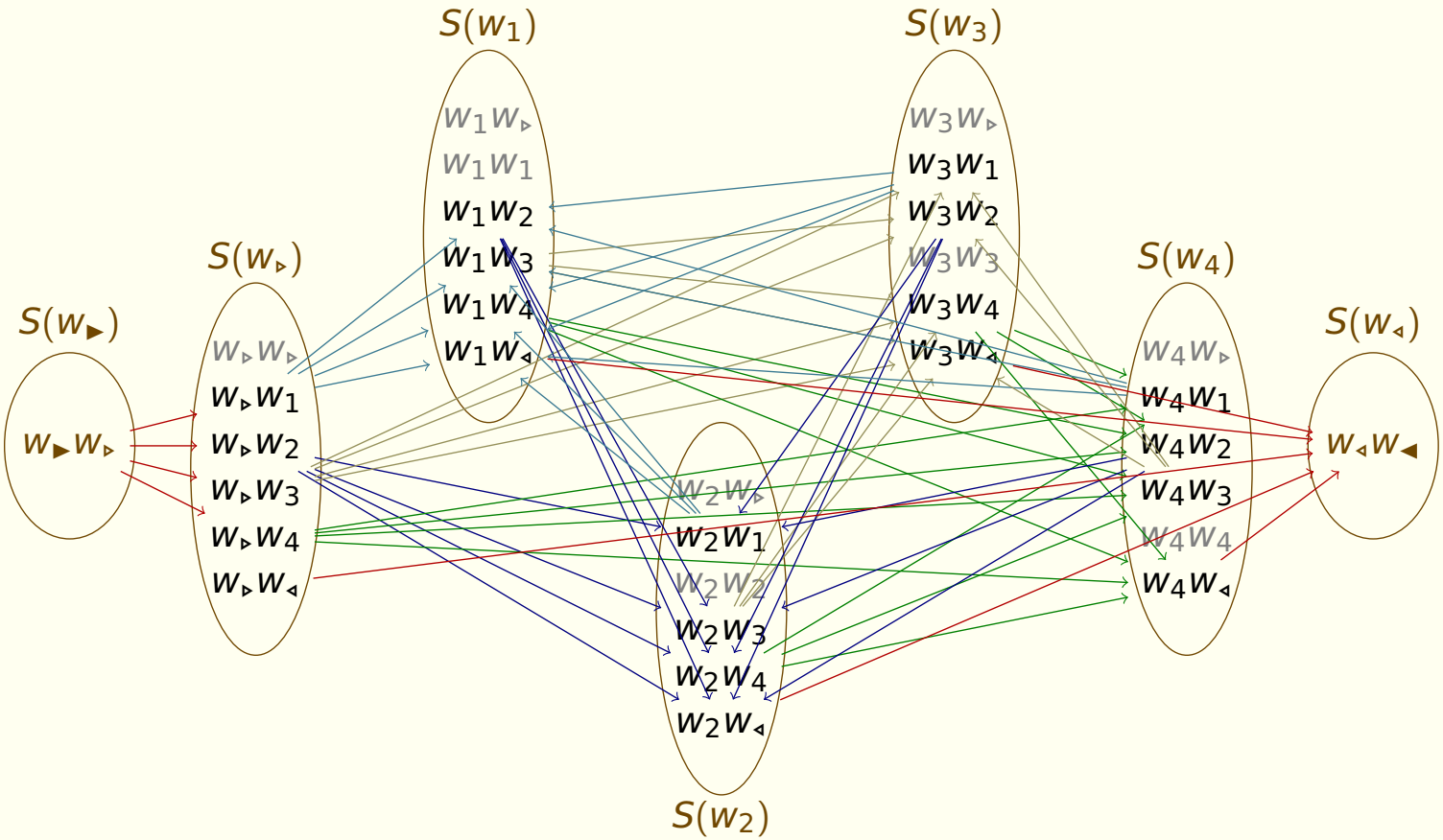
Word sequencing graph objective



Edge cost: $C[\langle w_i, w_j\rangle, \langle w_j, w_k\rangle] = -\log\left(P(w_k|w_i, w_j)\right)$

$$C_{total}(\tau) = \sum_{\forall w_i, w_j, w_k \in \mathcal{W}^3} \tau[\langle w_i, w_j\rangle, \langle w_j, w_k\rangle] \cdot C[\langle w_i, w_j\rangle, \langle w_j, w_k\rangle]$$

Word Use Constraint

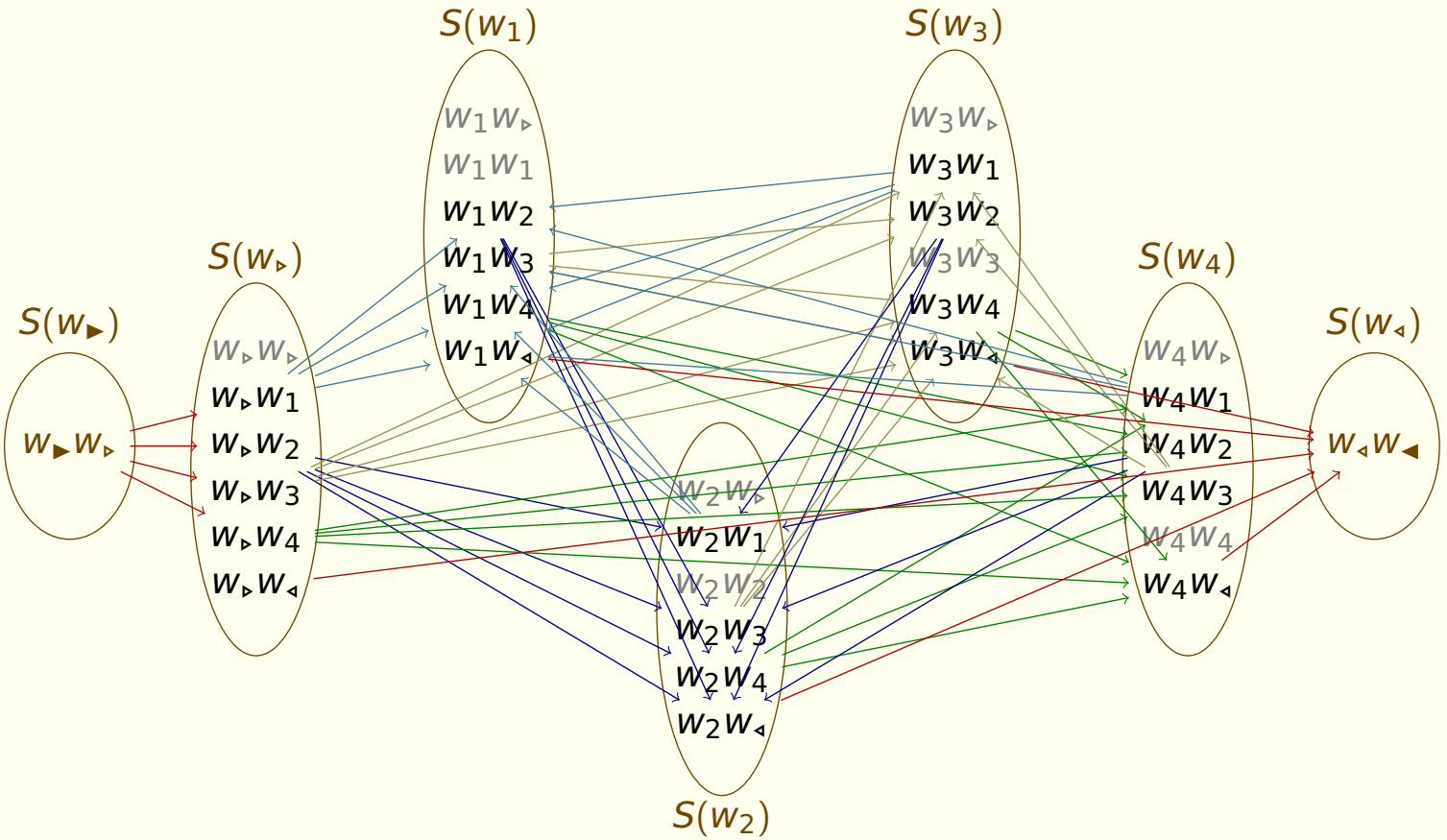


Districts: Every word must be used exactly once

$\forall w_i \in \mathcal{W} \backslash \{w_\blacktriangleright, w_\blacktriangleleft\}$:

$$\sum_{\forall \langle w_i, w_j \rangle \in S(w_i)} \sum_{\forall \langle w_h, w_i \rangle \in \mathcal{W}^2} \tau[\langle w_h, w_i \rangle, \langle w_i, w_j \rangle] = 1$$

Single Path Constraint

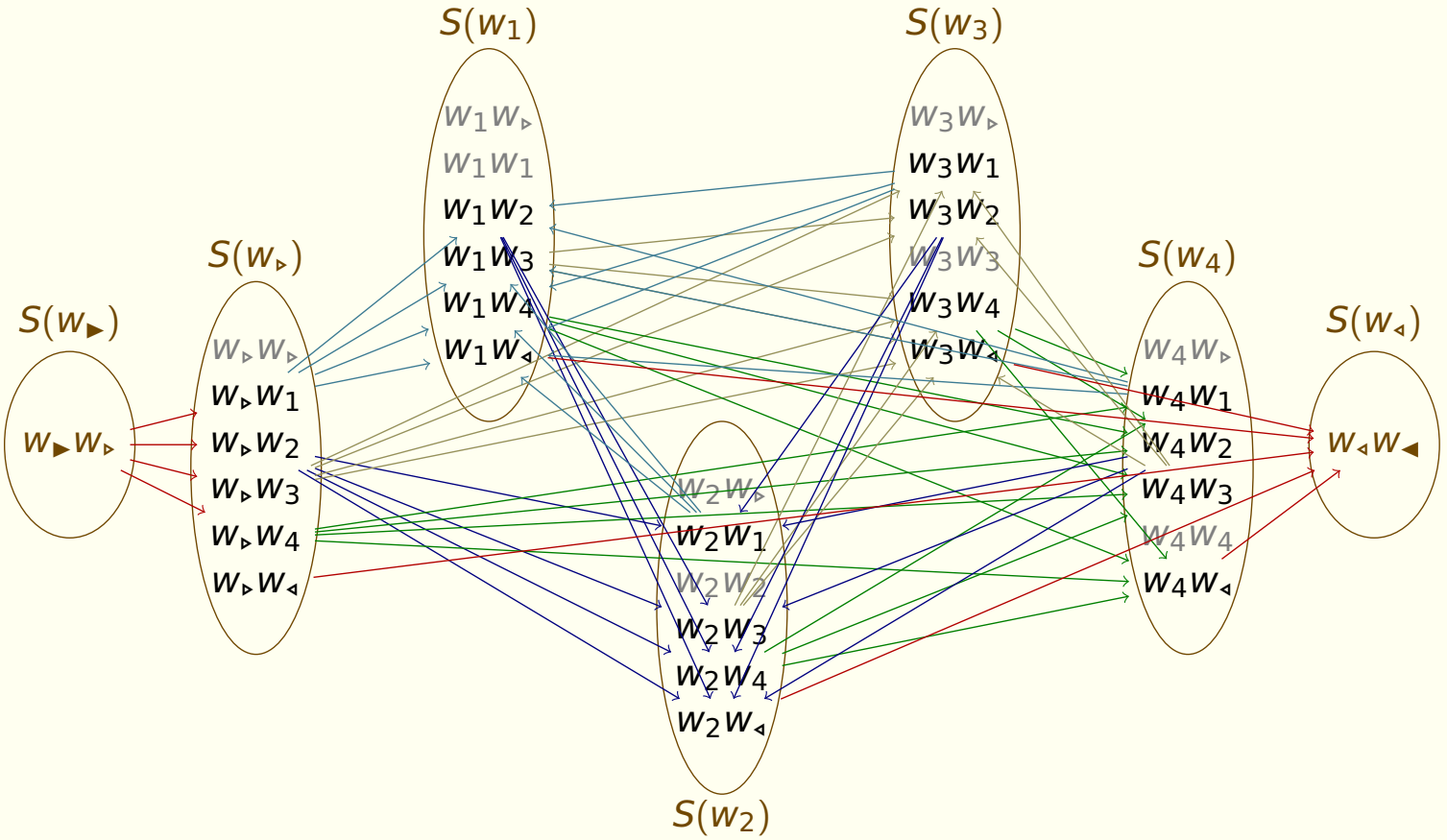


Single Path: every word node entered must be exited

$\forall \langle w_i, w_j \rangle \in \mathcal{W}^2 \setminus \{\langle w_\blacktriangleright, w_\triangleright \rangle, \langle w_\triangleleft, w_\blacktriangleleft \rangle\}$:

$$\sum_{\forall \langle w_a, w_b \rangle \in \mathcal{W}^2} \tau[\langle w_a, w_b \rangle, \langle w_i, w_j \rangle] = \sum_{\forall \langle w_c, w_d \rangle \in \mathcal{W}^2} \tau[\langle w_i, w_j \rangle, \langle w_c, w_d \rangle]$$

Single Path Constraints



Single Path: No Subtours

$T \subseteq \mathcal{W}^2$ including all nodes connected from $\langle w_\blacktriangleright, w_\triangleright \rangle$

$$\{ w_i \mid \forall \langle w_i, w_j \rangle \in T \} \cup \{ w_\blacktriangleleft \} = \mathcal{W}.$$

Results

Process	Perfect	Mean Precision	Mean Jaccard Index
Selection Only	75.6%	0.912	0.891

Process	Perfect	BLEU Score	Portion Feasible
Ordering Only	66.6%	0.806	99.6%
Full System	62.2%	0.745	93.7%

LSTM-VAE Examples

		Sel.	Ord.
Reference	i went to the kitchen .		
Full System	i went to the kitchen .	✓	✓
VAE Mean	i went to the kitchen .		
VAE Sample1	i went to my apartment .		
VAE Sample2	i looked around the room .		
VAE Sample3	i turned back to the table .		

Samuel R Bowman et al. “Generating Sentences from a Continuous Space”. In: *International Conference on Learning Representations (ICLR) Workshop* (2016).

DT-RAE Examples

		Sel.	Ord.
Reference	name this 1922 novel about leopold bloom written by james joyce .		
Full System	written novel by name james about leopold this bloom 1922 joyce .	✓	✖
DT-RAE Ref.	name this 1906 novel about gottlieb_fecknoe inspired by james_joyce		
DT-RAE Para.	what is this william golding novel by its written writer		

Mohit Iyyer, Jordan Boyd-Graber, and Hal Daumé III. "Generating Sentences from Semantic Vector Space Representations". In: *NIPS Workshop on Learning Semantics*. 2014.

DT-RAE Examples

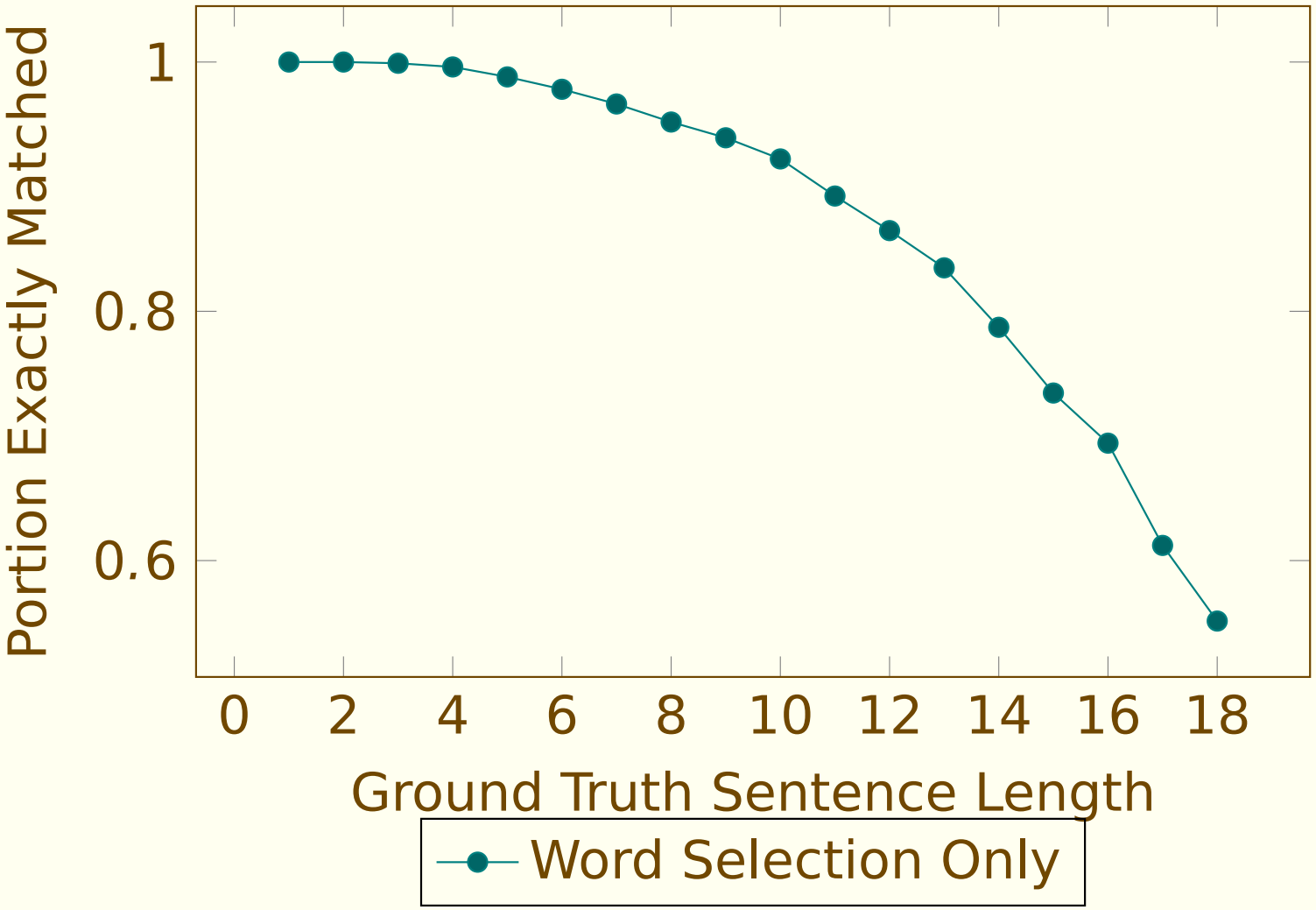
		Sel.	Ord.
Reference	this is the basis of a comedy of manners first performed in 1892 .		
Full System	this is the basis of a comedy of manners first performed in 1892 .	✓	✓
DT-RAE Ref.	another is the subject of this trilogy of romance most performed in 1874		
DT-RAE Para.	subject of drama from him about romance		

Mohit Iyyer, Jordan Boyd-Graber, and Hal Daumé III. "Generating Sentences from Semantic Vector Space Representations". In: *NIPS Workshop on Learning Semantics*. 2014.

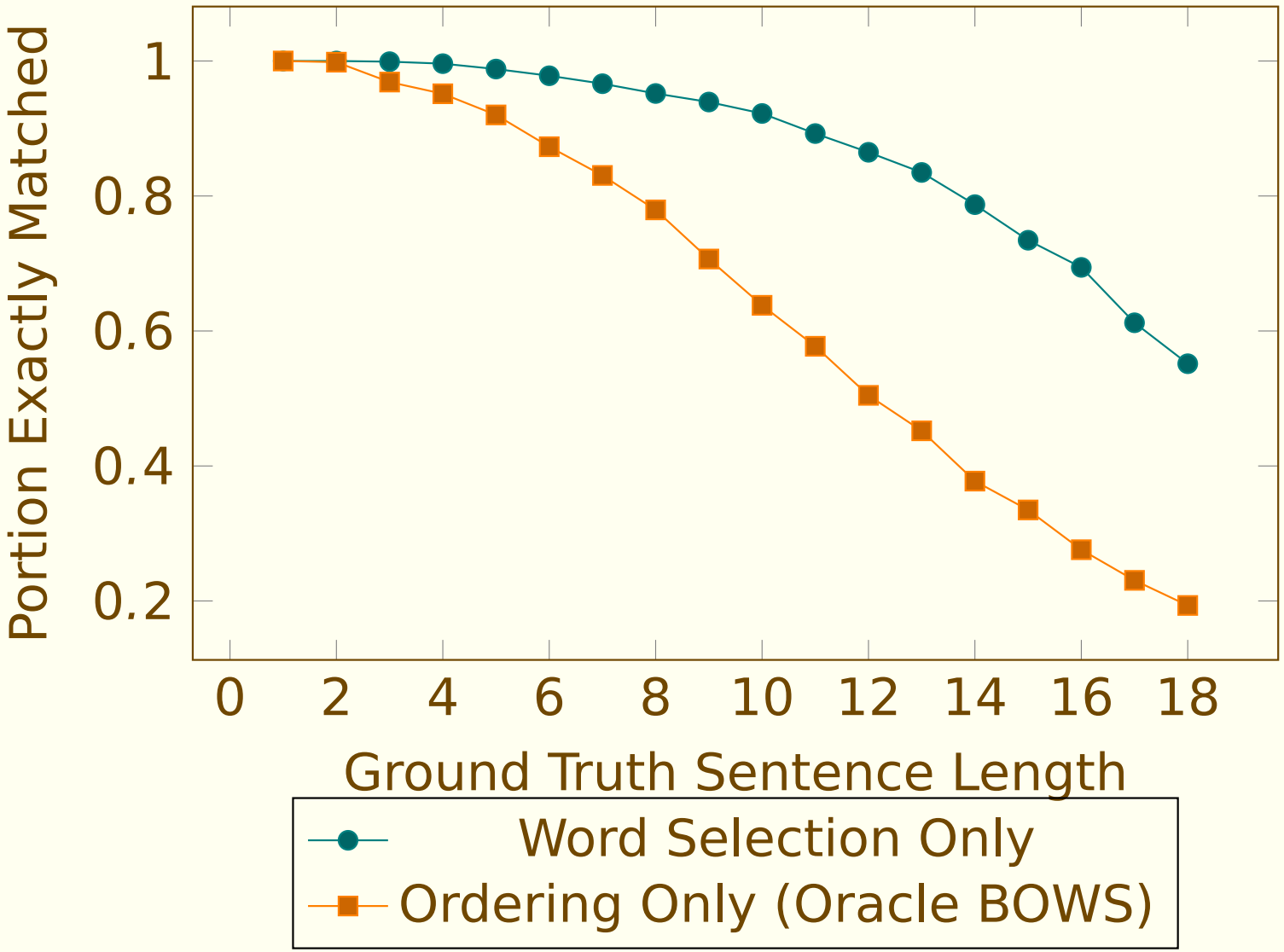
Ambiguous Examples

		Sel.	Ord.
Reference	please give me directions from Paris to London .		
Full System	please give me directions to London from Paris .	✓	✗

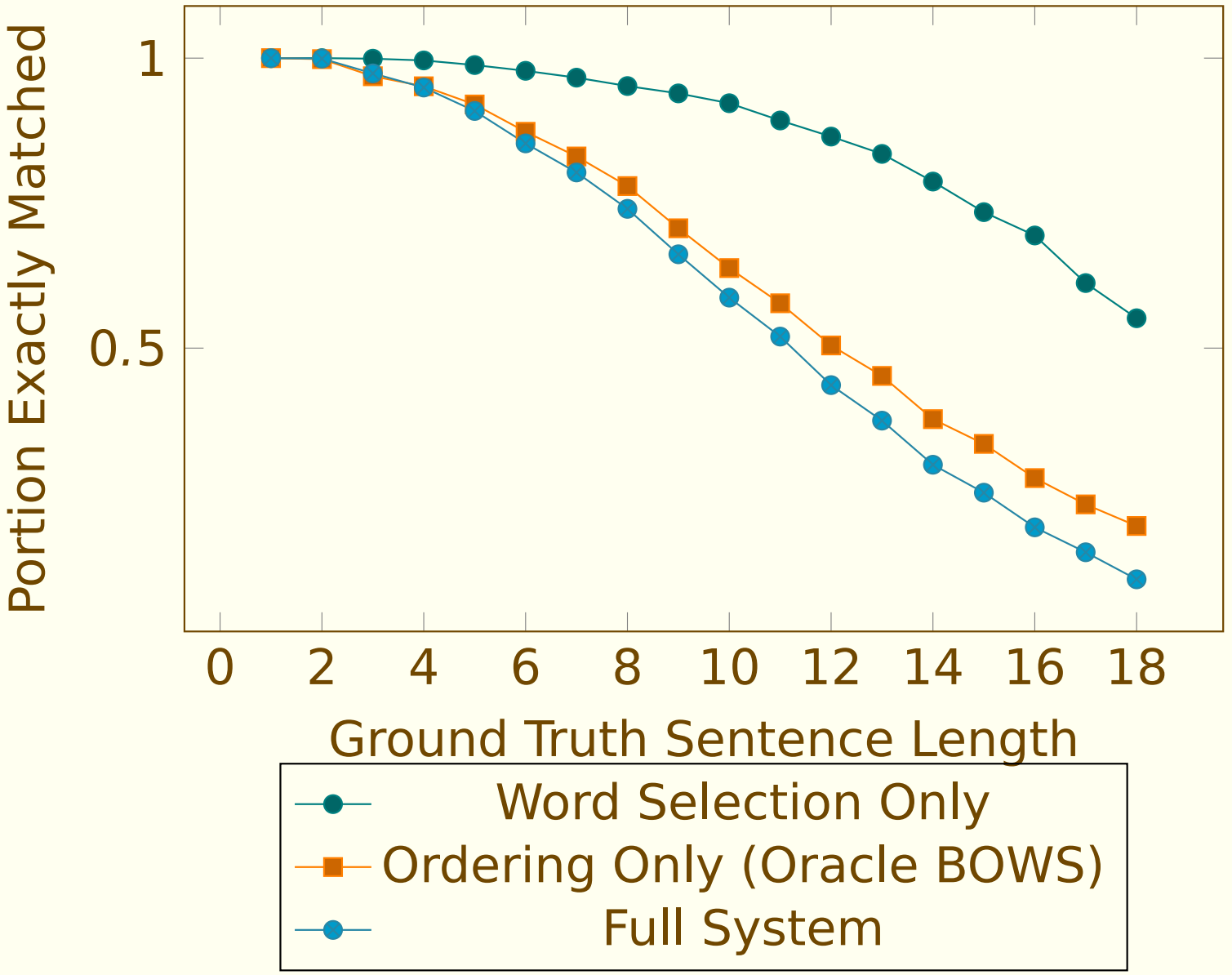
Per Length Results: Word Selection



Per Length Results: Ordering with Oracle BOW



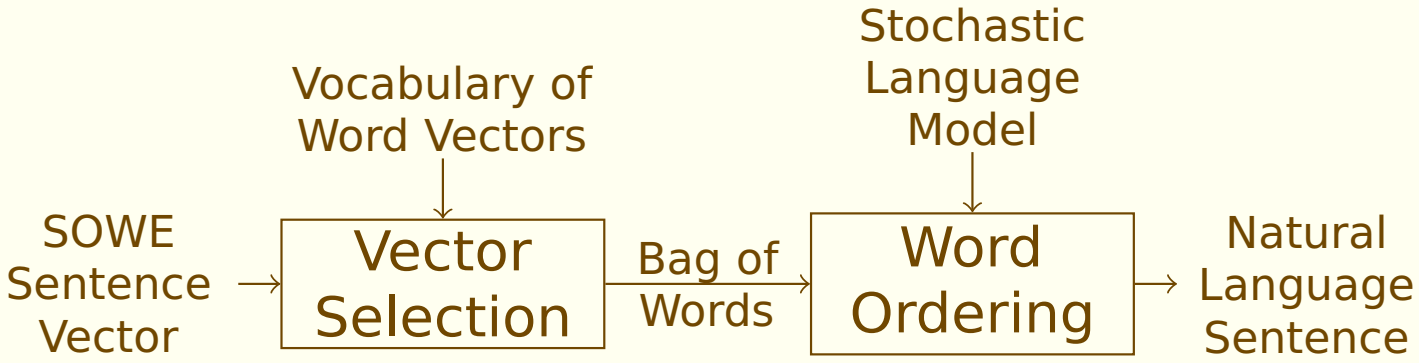
Per Length Results: Our complete system



Results summary

- 0.75 BLEU is very good, 62% perfect recreation is usable.
- This does come from most sentences being short though.
- This is the first work to present quantitative results for sentence generation based purely on a sentence vector representation.
- It does not produce paraphrases
- Its robustness against noise has not been investigated.

We broke the problem down into two subproblems.



Vector Selection: Select which word vectors go into the sum

Word Ordering: Find them most likely order of words

Conclusion: Split the sentence generation into selection then ordering.

- Selection: Greedy Method
  - Broad generalisation of Knapsack Problem
  - Input: vector
  - Apply greedy iterative method.
  - Output: BOW

Conclusion: Split the sentence generation into selection then ordering.

- Selection: Greedy Method
  - Broad generalisation of Knapsack Problem
  - Input: vector
  - Apply greedy iterative method.
  - Output: BOW

- Ordering:
  - Find most likely order using trigrams
  - Input: BOW
  - Rewrite MIP and solve
  - Output: Most likely order of words

Appendix

Experimental Setup: Language Modeling

- using subset of Books Corpus for Training and Testing
- Further subset to length 18 or less

Reinhard Kneser and Hermann Ney. "Improved backing-off for m-gram language modeling". In: *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*. Vol. 1. IEEE. 1995, pp. 181–184.

Experimental Setup: Language Modeling

- using subset of Books Corpus for Training and Testing
- Further subset to length 18 or less

- Language model trained on 6 million sentences.
- Use Kneser Ney Smoothing.

Reinhard Kneser and Hermann Ney. "Improved backing-off for m-gram language modeling". In: *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*. Vol. 1. IEEE. 1995, pp. 181–184.

Alternative notation for vector selection problem

Rather than writing:
Find the inclusion vector $\tilde{c} = [c_1, c_2, \ldots c_n] \in \mathbb{N}_0^n$ to minimise $d(\tilde{s}, \sum_{\tilde{x}_j \in \mathcal{V}} c_j \tilde{x}_j)$

Write:
Find the bag of vectors $\mathcal{B}$ (a multi-subset of $\mathcal{V}$), such that $\Sigma(\mathcal{B}) = \sum_{\tilde{x}_a \in \mathcal{B}} \tilde{x}_a$ we have $\min d(w_\blacktriangleright, \Sigma(\mathcal{B}))$

Greedy Addition: where you add the best vector to your current bag, and repeat.

Find the bag of vectors $\mathcal{B}$ (a multi-subset of $\mathcal{V}$), such that $\Sigma(\mathcal{B}) = \sum_{\tilde{x}_a \in \mathcal{B}} \tilde{x}_a$ we have $\min d(w_\blacktriangleright, \Sigma(\mathcal{B}))$

1. For each vector $\tilde{x}_j$ in the vocabulary consider $d(w_\blacktriangleright, \Sigma(\mathcal{B}) + \tilde{x}_j)$
2. Add the vector that gets closest the bag. $\mathcal{B} \leftarrow \mathcal{B} \cup \{\tilde{x}_\star\}$
   - unless adding nothing would be better – then terminate
3. Repeat

A 1 dimensional example of greedy addition

Find the bag of vectors $\mathcal{B}$ (a multi-subset of $\mathcal{V}$), such that $\Sigma(\mathcal{B}) = \sum_{\tilde{x}_a \in \mathcal{B}} \tilde{x}_a$ we have $\min d(w_\blacktriangleright, \Sigma(\mathcal{B}))$

Consider $\mathcal{V} = \{24, 25, 100\}$ $\quad w_\blacktriangleright = 148 \quad d(x, y) = |x - y|$

1. $\mathcal{B} = []$ $\qquad d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - 0| = 149$

A 1 dimensional example of greedy addition

Find the bag of vectors $\mathcal{B}$ (a multi-subset of $\mathcal{V}$), such that $\Sigma(\mathcal{B}) = \sum_{\tilde{x}_a \in \mathcal{B}} \tilde{x}_a$ we have $\min d(w_\blacktriangleright, \Sigma(\mathcal{B}))$

Consider $\mathcal{V} = \{24, 25, 100\}$ $\quad w_\blacktriangleright = 148 \quad d(x, y) = |x - y|$

1. $\mathcal{B} = []$ $\qquad d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - 0| = 149$
2. $\mathcal{B} = [100]$ $\qquad d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - 100| = 48$

A 1 dimensional example of greedy addition

Find the bag of vectors $\mathcal{B}$ (a multi-subset of $\mathcal{V}$), such that $\Sigma(\mathcal{B}) = \sum_{\tilde{x}_a \in \mathcal{B}} \tilde{x}_a$ we have $\min d(w_\blacktriangleright, \Sigma(\mathcal{B}))$

Consider $\mathcal{V} = \{24, 25, 100\}$ $w_\blacktriangleright = 148$ $d(x, y) = |x - y|$

1. $\mathcal{B} = []$ $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - 0| = 149$
2. $\mathcal{B} = [100]$ $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - 100| = 48$
3. $\mathcal{B} = [100, 25]$ $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - (100 + 25)| = 23$

A 1 dimensional example of greedy addition

Find the bag of vectors $\mathcal{B}$ (a multi-subset of $\mathcal{V}$), such that $\Sigma(\mathcal{B}) = \sum_{\tilde{x}_a \in \mathcal{B}} \tilde{x}_a$ we have $\min d(w_\blacktriangleright, \Sigma(\mathcal{B}))$

Consider $\mathcal{V} = \{24, 25, 100\}$ $w_\blacktriangleright = 148$ $d(x, y) = |x - y|$

1. $\mathcal{B} = []$ $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - 0| = 149$
2. $\mathcal{B} = [100]$ $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - 100| = 48$
3. $\mathcal{B} = [100, 25]$ $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - (100 + 25)| = 23$
4. $\mathcal{B} = [100, 25, 24]$
   $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - (100 + 25 + 24)| = 1$

A 1 dimensional example of greedy addition

Find the bag of vectors $\mathcal{B}$ (a multi-subset of $\mathcal{V}$), such that $\Sigma(\mathcal{B}) = \sum_{\tilde{x}_a \in \mathcal{B}} \tilde{x}_a$ we have $\min d(w_\blacktriangleright, \Sigma(\mathcal{B}))$

Consider $\mathcal{V} = \{24, 25, 100\}$ $w_\blacktriangleright = 148$ $d(x, y) = |x - y|$

1. $\mathcal{B} = []$ $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - 0| = 149$
2. $\mathcal{B} = [100]$ $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - 100| = 48$
3. $\mathcal{B} = [100, 25]$ $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - (100 + 25)| = 23$
4. $\mathcal{B} = [100, 25, 24]$
   $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - (100 + 25 + 24)| = 1$
5. $\mathcal{B} = [100, 25, 24]$ No improvement possible
   $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = 1$

Fell for greedy trap

1-Substitution: Lessen the greed by reconsidering past choices

Find the bag of vectors $\mathcal{B}$ (a multi-subset of $\mathcal{V}$), such that $\Sigma(\mathcal{B}) = \sum_{\tilde{x}_a \in \mathcal{B}} \tilde{x}_a$ we have $\min d(w_\blacktriangleright, \Sigma(\mathcal{B}))$

1. Consider each word vector in the current bag $\tilde{x}_a \in \mathcal{B}$
2. Would deleting it improve the score?
   $d(w_\blacktriangleright, \Sigma(\mathcal{B}) - \tilde{x}_a) < d(w_\blacktriangleright, \Sigma(\mathcal{B}))$?
3. Can it be swapped for another word to improve the score? $\exists \tilde{x}_b \in \mathcal{V}$ such that
   $d(w_\blacktriangleright, \Sigma(\mathcal{B}) - \tilde{x}_a + \tilde{x}_b)) < d(w_\blacktriangleright, \Sigma(\mathcal{B}))$?

A 1 dimensional example of 1-substitution

Find the bag of vectors $\mathcal{B}$ (a multi-subset of $\mathcal{V}$), such that $\Sigma(\mathcal{B}) = \sum_{\tilde{x}_a \in \mathcal{B}} \tilde{x}_a$ we have $\min d(w_\blacktriangleright, \Sigma(\mathcal{B}))$

Consider $\mathcal{V} = \{24, 25, 100\}$ $\quad w_\blacktriangleright = 148$ $\quad d(x, y) = |x - y|$

1. $\mathcal{B} = [100, 25, 24]$
   $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - (100 + 25 + 24)| = 1$

A 1 dimensional example of 1-substitution

Find the bag of vectors $\mathcal{B}$ (a multi-subset of $\mathcal{V}$), such that $\Sigma(\mathcal{B}) = \sum_{\tilde{x}_a \in \mathcal{B}} \tilde{x}_a$ we have $\min d(w_\blacktriangleright, \Sigma(\mathcal{B}))$

Consider $\mathcal{V} = \{24, 25, 100\}$ $\quad w_\blacktriangleright = 148 \quad d(x, y) = |x - y|$

1. $\mathcal{B} = [100, 25, 24]$
   $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - (100 + 25 + 24)| = 1$
2. $\mathcal{B} = [100, 24, 24]$
   $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - (100 + 24 + 24)| = 0$

A 1 dimensional example of 1-substitution

Find the bag of vectors $\mathcal{B}$ (a multi-subset of $\mathcal{V}$), such that $\Sigma(\mathcal{B}) = \sum_{\tilde{x}_a \in \mathcal{B}} \tilde{x}_a$ we have $\min d(w_\blacktriangleright, \Sigma(\mathcal{B}))$

Consider $\mathcal{V} = \{24, 25, 100\}$ $w_\blacktriangleright = 148$ $d(x, y) = |x - y|$

1. $\mathcal{B} = [100, 25, 24]$
   $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - (100 + 25 + 24)| = 1$
2. $\mathcal{B} = [100, 24, 24]$
   $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = |148 - (100 + 24 + 24)| = 0$
3. $\mathcal{B} = [100, 24, 24]$ Perfect $d(w_\blacktriangleright, \Sigma(\mathcal{B})) = 0$

Fixed, but there are deeper greed traps, that can be constructed.

Experimental Setup: Pre-process corpora to only use known words.

- For word embeddings, we use pretrained GloVe

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. 2014, pp. 1532–1543.

Experimental Setup: Pre-process corpora to only use known words.

- For word embeddings, we use pretrained GloVe
- Restrict Vector vocab to only words used in corpora
- Pre-process Corpora to remove sentences with words not found in vocabulary.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. 2014, pp. 1532–1543.

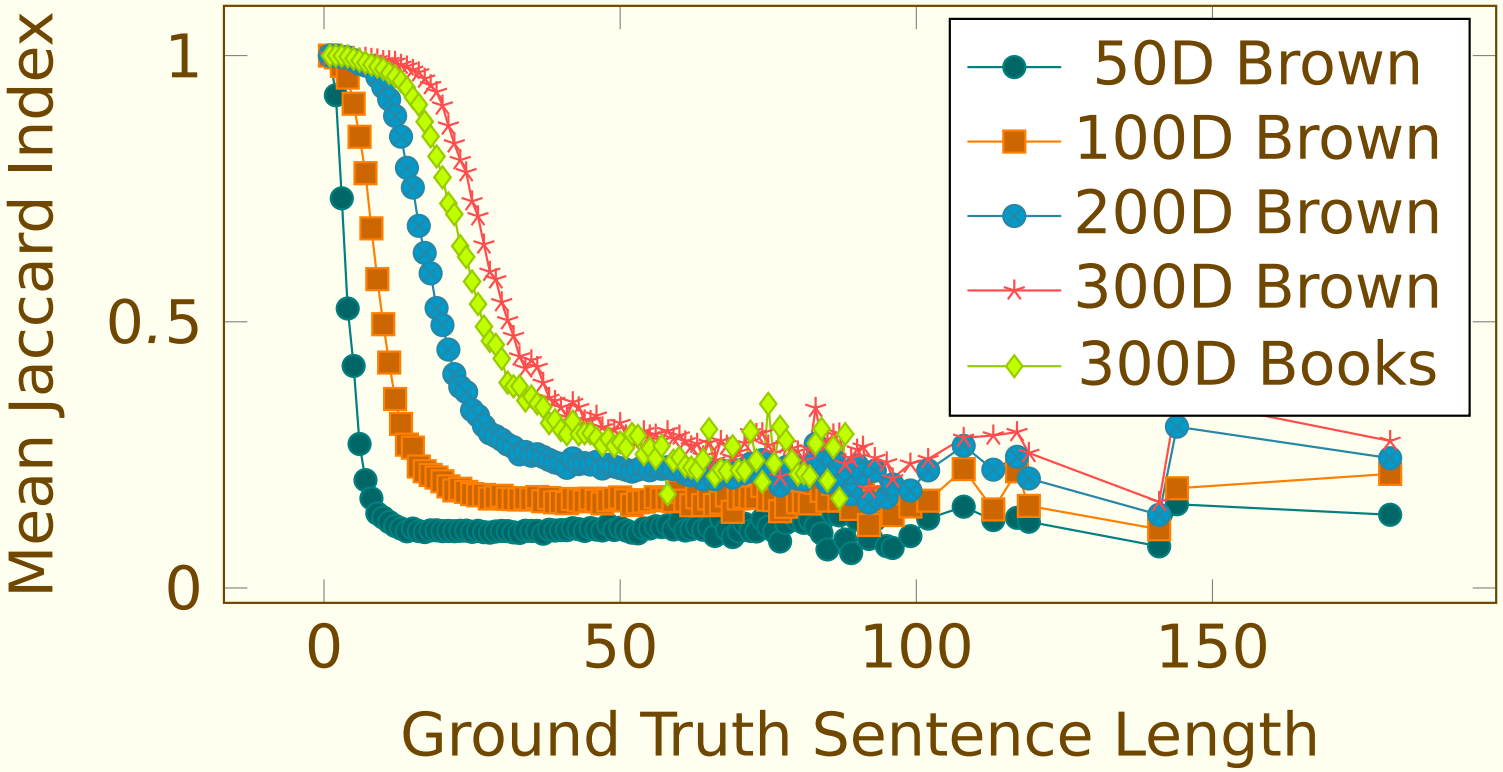
Experimental Setup: we used the Books Corpus

- 178,694 unique words
- 66,464 sentences
- Sentence Length Q3: 17 words
- 11,038 unpublished novels, we use just a small random subset

---

Yukun Zhu et al. “Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books”. In: *arXiv preprint arXiv:1506.06724*. 2015.

Word Selection Results



DT-RAE Examples

		Sel.	Ord.
Reference	name this 1922 novel about leopold bloom written by james joyce .		
Full System	written novel by name james about leopold this bloom 1922 joyce .	✓	✖
DT-RAE Ref.	name this 1906 novel about gottlieb_fecknoe inspired by james_joyce		
DT-RAE Para.	what is this william golding novel by its written writer		

Mohit Iyyer, Jordan Boyd-Graber, and Hal Daumé III. "Generating Sentences from Semantic Vector Space Representations". In: *NIPS Workshop on Learning Semantics*. 2014.

DT-RAE Examples

		Sel.	Ord.
Reference	ralph waldo emerson dismissed this poet as the jingle man and james russell lowell called him three-fifths genius and two-fifths sheer fudge .		
Full System	him “ james great as emerson genius ralph the lowell and sheer waldo three-fifths man fudge dismissed jingle russell two-fifths and gwalchmai 2009 vice-versa	✗	✖
DT-RAE Ref.	prominent called 21.25 explained henry_david_thoreau rejected this author like the tsar boat and imbalance created known good writing and his own death		
DT-RAE Para.	henry_david_thoreau rejected him through their stories to go money well inspired stories to write as her writing		

Mohit Iyyer, Jordan Boyd-Graber, and Hal Daumé III. “Generating Sen-

DT-RAE Examples

		Sel.	Ord.
Reference	this is the basis of a comedy of manners first performed in 1892 .		
Full System	this is the basis of a comedy of manners first performed in 1892 .	✓	✓
DT-RAE Ref.	another is the subject of this trilogy of romance most performed in 1874		
DT-RAE Para.	subject of drama from him about romance		

Mohit Iyyer, Jordan Boyd-Graber, and Hal Daumé III. "Generating Sentences from Semantic Vector Space Representations". In: *NIPS Workshop on Learning Semantics*. 2014.

DT-RAE Examples

		Sel.	Ord.
Reference	in a third novel a sailor abandons the patna and meets marlow who in another novel meets kurtz in the congo .		
Full System	kurtz and another meets sailor meets the marlow who abandons a third novel in a novel in the congo in patna .	✓	✗
DT-RAE Ref.	during the short book the lady seduces the family and meets cousin he in a novel dies sister from the mr.		
DT-RAE Para.	during book of its author young lady seduces the family to marry old suicide while i marries himself in marriage		

Mohit Iyyer, Jordan Boyd-Graber, and Hal Daumé III. "Generating Sentences from Semantic Vector Space Representations". In: *NIPS Workshop on Learning Semantics*. 2014.

DT-RAE Examples

		Sel.	Ord.
Reference	thus she leaves her husband and child for aleksei vronsky but all ends sadly when she leaps in front of a train .		
Full System	she her all when child for leaves front but and train ends husband aleksei leaps of vronsky in a sadly micro-history thus , she the	✗	✖
DT-RAE Ref.	however she leaves her sister and daughter from former fiancé and she ends unfortunately when narrator drives into life of a house		
DT-RAE Para.	leaves the sister of man in this novel		

Mohit Iyyer, Jordan Boyd-Graber, and Hal Daumé III. "Generating Sentences from Semantic Vector Space Representations". In: *NIPS Workshop on Learning Semantics*. 2014.

LSTM-VAE Examples

		Sel.	Ord.
Reference	we looked out at the setting sun .		
Full System	we looked out at the setting sun .	✓	✓
VAE Mean	they were laughing at the same time .		
VAE Sample1	ill see you in the early morning .		
VAE Sample2	i looked up at the blue sky .		
VAE Sample3	it was down on the dance floor .		

Samuel R Bowman et al. “Generating Sentences from a Continuous Space”. In: *International Conference on Learning Representations (ICLR) Workshop* (2016).

LSTM-VAE Examples

		Sel.	Ord.
Reference	i went to the kitchen .		
Full System	i went to the kitchen .	✓	✓
VAE Mean	i went to the kitchen .		
VAE Sample1	i went to my apartment .		
VAE Sample2	i looked around the room .		
VAE Sample3	i turned back to the table .		

Samuel R Bowman et al. “Generating Sentences from a Continuous Space”. In: *International Conference on Learning Representations (ICLR) Workshop* (2016).

LSTM-VAE Examples

		Sel.	Ord.
Reference	how are you doing ?		
Full System	how ’re do well ?	✗	✗
VAE Mean	what are you doing ?		
VAE Sample1	“ are you sure ?		
VAE Sample2	what are you doing, ?		
VAE Sample3	what are you doing ?		

Samuel R Bowman et al. “Generating Sentences from a Continuous Space”. In: *International Conference on Learning Representations (ICLR) Workshop* (2016).

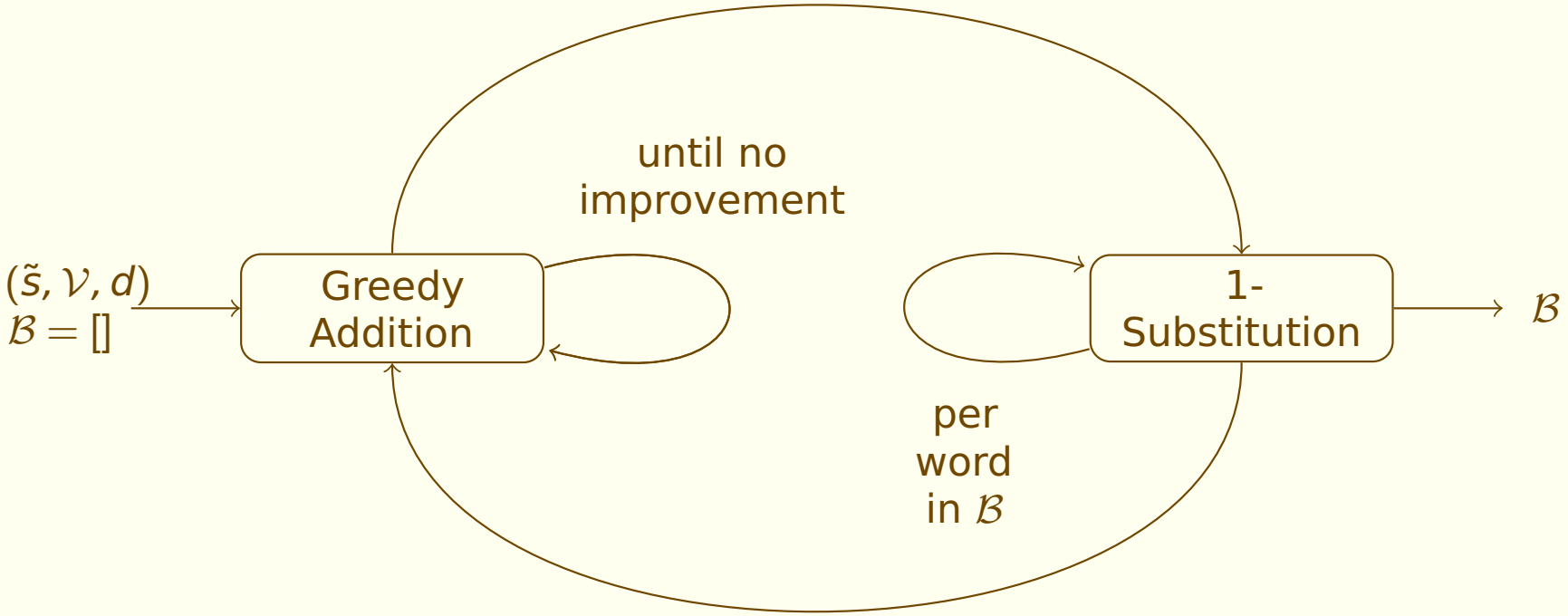
Ambiguous Examples

		Sel.	Ord.
Reference	it was the worst of times , it was the best of times .		
Full System	it was the best of times , it was the worst of times .	✓	✗

Ambiguous Examples

		Sel.	Ord.
Reference	please give me directions from Paris to London .		
Full System	please give me directions to London from Paris .	✓	✗

Run until convergence



- Greedy Addition: add word to bag that gets us closest
- 1-Substitution: Reconsider past choices

Lyndon White et al. “Generating Bags of Words from the Sums of their Word Embeddings”. In: *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. 2016.