Lecture 8: Word Sense Disambiguation

Xu Ruifeng

Harbin Institute of Technology, Shenzhen

Last Time

- Word Meanings
- Computer Expression of Word Meanings
- English Semantic Resources
  - WordNet
- Chinese Resources
  - CiLin
  - HowNet
  - Chinese Concept Dictionary: CCD

Today’s Class

- Word Sense Disambiguation
- Background and Restrictions
- Automatic Word Sense Disambiguation
  - Knowledge-based Approach
  - Machine Learning based Approach
    - Supervised Method
    - Semi-supervised Method
    - Unsupervised Method
  - Hybrid Approach
- Evaluations

Word Sense Disambiguition

In our house, everybody has a career and none of them includes washing dishes.

I'm looking for a restaurant that serves vegetarian dishes.

- Many words have multiple senses
- Task: determine which of various senses of a word are invoked in context

Examples (Yarowsky, 1995)

plant	living/factory
tank	vehicle/container
poach	steal/boil
palm	tree/hand
bass	fish/music
motion	legal/physical
crane	bird/machine

Harder Cases

- WordNet： Senses of “Line”

(1) a formation of people or things one behind another

(2) length (straight or curved) without breadth or thickness; the trace of a moving point

(3) space for one line of print (one column wide and 1/14 inch deep) used to measure advertising;

(4) a fortified position (especially one marking the most forward position of troops);

(5) a slight depression in the smoothness of a surface;

(6) something (as a cord or rope) that is long and thin and flexible;

(7) the methodical process of logical reasoning;

(8) the road consisting of railroad track and roadbed;

WSD: Motivation

- One of the central challenges in NLP.
- Ubiquitous across all languages.
- Needed in:
  - Machine Translation: For correct lexical choice.
  - Information Retrieval: Resolving ambiguity in queries.
  - Information Extraction: For accurate analysis of text.
- Computationally determining which sense of a word is activated by its use in a particular context.

WSD: Motivation

- Question answering
- Natural language generation
- Language modeling
- Automatic essay grading
- Plagiarism detection
- Document clustering

Automatic WSD Approaches

- Knowledge Based Approach
  - WSD using Selectional Preferences (or restrictions)
  - Overlap Based Approaches
- Machine Learning Based Approach
  - Supervised Approaches
  - Semi-supervised Algorithms
  - Unsupervised Algorithms
- Hybrid Approach
- Neural Approach

Knowledge-based Approaches: Restrictions

- Constraints imposed by syntactic dependencies
  - I love washing dishes
  - I love spicy dishes
- Selectional restrictions may be too weak
  - I love this dish
- Early work: semantic networks, frames, logical reasoning and "expert systems" (Hirst, 1988)
- Brown et al. (1991), Resnik (1993)
  - Non-standard indicators: tense, adjacent words for collocations (*mace spray*; *mace and parliament*)

Overlap based Approach

- Require a ***Machine Readable Dictionary***
- Find the overlap between the features of different senses of target word (sense bag) and features of the words in its context (context bag).
- These features could be sense definitions, example sentences, hypernyms etc.
- The features could also be given weights
- The sense which has the maximum overlap is selected as the contextually appropriate one

LESK'S Algorithm

Sense Bag: *contains the words in the definition of a candidate sense of the ambiguous word.*

Context Bag: *contains the words in the definition of each sense of each context word.*

E.g. "On burning ***coal*** we get ***ash***."

- Ash
- Sense 1
  Trees of the olive family with pinnate leaves, thin furrowed bark and gray branches.
- Sense 2
  The ***solid*** residue left when ***combustible*** material is thoroughly ***burn***ed or oxidized.
- Sense 3
  To convert into ash

- Coal
- Sense 1
  A piece of glowing carbon or ***burn***t wood.
- Sense 2
  charcoal.
- Sense 3
  A black ***solid combustible*** substance formed by the partial decomposition of vegetable matter without free access to air and under the influence of moisture and often increased pressure and temperature that is widely used as a fuel for ***burn***ing

- In this case Sense 2 of ash is selected

WALKER'S Thesaurus Based Algorithm

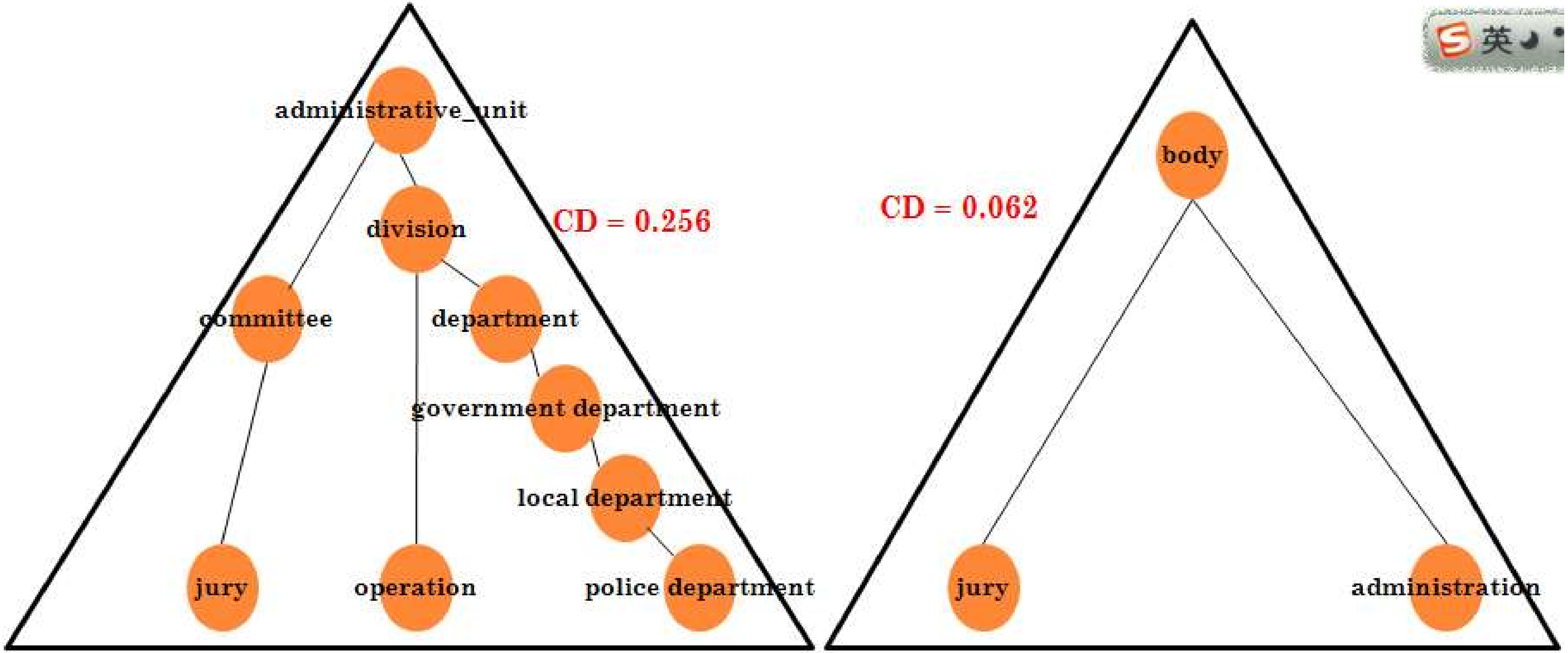
- ***Step 1*: For each sense of the target word find the thesaurus category to which that sense belongs.**
- ***Step 2*: Calculate the score for each sense by using the context words. A context words will add 1 to the score of the sense if the thesaurus category of the word matches that of the sense.**
  - *E.g. The money in this **bank** fetches an interest of 8% per annum*
  - Target word: ***bank***
  - Clue words from the context: ***money***, ***interest***, ***annum***, ***fetch***

	Sense1: Finance	Sense2: Location
Money	+1	0
Interest	+1	0
Fetch	0	0
Annum	+1	0
Total	3	0

Context words add 1 to the sense when the topic of the word matches that of the sense

Conceptual Density based WSD

- Select a sense based on the relatedness of that word-sense to the context.
- Relatedness is measured in terms of conceptual distance
  - i.e. how close the concept represented by the word and the concept represented by its context words are
- This approach uses WordNet for finding the conceptual distance.
- Smaller the conceptual distance higher will be the conceptual density.



The jury(2) praised the administration(3) and **operation** (8) of Atlanta Police Department(1)

Step 1: Make a lattice of the nouns in the context, their senses and hypernyms.

Step 2: Compute the conceptual density of resultant concepts

Step 3: The concept with highest CD is selected.

Step 4: Select the senses below the selected concept as the correct sense for the respective words.

Knowledge-based Approach: Summary

- Using Selectional Restrictions
  - Needs exhaustive Knowledge Base
- Overlap based approaches
  - Dictionary definitions are generally very small
  - Dictionary entries rarely take into account the distributional constraints of different word senses (e.g. cigarette / ash)
  - Suffer from the problem of sparse match
  - Proper nouns are not present in a MRD
    - E.g. “Jordan” is a strong indicator of the category “sports - basketball”

Knowledge-based Approach: Summary

- We don't have a thesaurus for every language
- Even if we do, they have problems with recall
  - Many words are missing
  - Most (if not all) phrases are missing
  - Some connections between senses are missing
  - Not applicable well to verbs and adjectives
    - Adjectives and verbs have less structured hyponymy relations

Machine Learning based WSD

- Machine learning algorithm is applied to process WSD as a classification
- Machine Learning based Approaches
  - Supervised Approach
  - Semi-supervised Approach
  - Unsupervised Approach

Two Assumptions (Yarowsky 1995)

- **One Sense Per Collocation**
  - nearby words provide strong and consistent clues to the sense of a target word, conditional on relative distance, order and syntactic relationship
  - 95-96% applicable
- **One Sense Per Discourse**
  - The sense of a target word is highly consistent within any given document.
  - True for topic dependent words
  - Not true for verbs
  - Krovetz (1998): not true with respect to fine-grained senses: (e.g., language/people)

Intuition of distributional word similarity

- Nida example:

  A bottle of ***tesgüinois*** on the table!
  Everybody likes ***tesgüino***!
  Tesgüino makes you drunk!
  We make ***tesgüino*** out of corn.

- From context words, humans guess ***tesgüino*** means an alcoholic beverage like ***beer***.
- Intuition for algorithm:
  - Two words are similar if they have similar word contexts.

Reminder: Term document matrix

- Each cell: count of term *t* in a document *d*: $tf_{t,d}$:
  - Each document is a count vector in $\mathbb{N}^v$: a column below

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0

Reminder: Term document matrix

- Two documents are similar if their vectors are similar

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0

Reminder: Term document matrix

- Each word is a $\boxed{\text{count vector}}$ in $\mathbb{N}^D$: a row below

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0

Reminder: Term document matrix

- Two **words** are similar if their vectors are similar

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0

Reminder: Term document matrix

- Instead of using entire documents, use smaller contexts
  - Paragraph
  - Window of 10 words
- A word is now defined by a vector over counts of context words
- For the term-document matrix
  - We used ***tf-idf*** instead of raw term counts
  - ***Positive Pointwise Mutual Information (PPMI)*** is common (***PPMI*** Replace all ***PMI*** values less than 0 with zero)

Supervised Methods for WSD

- (Firth)"You shall know the word by the company it keeps"
- Supervised sense disambiguation is very successful
- However, it requires a lot of data

- A supervised method : Decision List

WSD: Naïve Bayes

$$s\hat{} = \operatorname{argmax}_{s\ \varepsilon\ senses} Pr(s|V_w)$$

where V_w is the feature vector.

Apply Bayes rule:

$$Pr(s|V_w)=Pr(s).Pr(V_w|s)/Pr(V_w)$$

$Pr(V_w|s)$ can be approximated by independence assumption:

$$Pr(V_w|s) = Pr(V_w^1|s)....Pr(V_w^n|s,V_w^1,..,V_w^{n-1})$$

$$= \Pi_{i=1}^n Pr(V_w^i|s)$$

$$s\hat{} = \operatorname{argmax}_{s\hat{I}senses} Pr(s).\Pi_{i=1}^n Pr(V_w^i|s)$$

Contextual Features in WSD

- Word found in +/⁻ k word window
- Word immediately to the right (+1 W)
- Word immediately to the left (-1 W)
- Pairs of words at offsets -2 and -1
- Pair of words at offsets -1 and +1
- Pair of words at offsets +1 and +2
- Part of speech of contextual words
- ……
- Some features are represented by their classes (WEEKDAY, MONTH)

Example

The ocean reflects the color of the sky, but even on cloudless days the color of the ocean is not a consistent blue. Phytoplankton, microscopic plant life that floats freely in the lighted surface waters, may alter the color of the water. When a great number of organisms are concentrated in an area, ...

w_{-1} = microscopic t_1=JJ

w_{+1} = life t_{+1}=NN

w_{-2},w_{-1}=(Phytoplankton, microscopic) ...

w_{-1},w_{+1}= (microscopic, life)

word-within-k=ocean

word-within-k=reflects

...

Decision Lists

- For each feature, we get an estimate of conditional probability of $sense_1$ and $sense_2$
- Consider the feature w_{+1}=life:

  Count($plant_1$, w_{+1} =life) =100

  Count($plant_2$, w_{+1} =life)=1
- Maximum-likelihood estimate

  $P(plant_1 | w_{+1} = life) = 100/101$
- Problem: sparse counts
- Use smoothing techniques

Creating Decision Lists

- For each feature, find

  sense(feature)=argmax_{sense} P(sense|feature)

  e.g., sense(w_{+1} =life)=$sense_1$
- Create a rule feature→ sense(feature)with

  weight P(sense(feature)|feature)

Rule	Weight
$w_{+1} = life \rightarrow plant_1$	0.99
$w_{+1} = work \rightarrow plant_2$	0.93

- Create a list of rules sorted by strength

Rule	Weight
$w_{+1} = life \rightarrow plant_1$	0.99
$w_{-1} = modern \rightarrow plant_2$	0.98
$w_{+1} = work \rightarrow plant_2$	0.975
word-within-k$= life \rightarrow plant_1$	0.95
$w_{-1} = assembly \rightarrow plant_2$	0.94

- To apply the decision list: take the first rule in the list which applies to an example

Applying Decision Lists

The ocean reflects the color of the sky, but even on cloudless days the color of the ocean is not a consistent blue. Phytoplankton, microscopic plant life that floats freely in the lighted surface waters, may alter the color of the water. When a great number of organisms are concentrated in an area, ...

Feature	Sense	Strength
w_{-1} = microscopic	1	0.95
w_{+1} = life	1	0.99
w_{-2}, w_{-1} =	N/A	
word-within-k=reflects	2	0.65
...		

- N/A → feature has not seen in training data

  w_{+1} = life Sense_1 is chosen

Experimental Results

- (Yarowsky, 1995)
- Accuracy of 95% on binary WSD

plant	living/factory
tank	vehicle/container
poach	steal/boil
palm	tree/hand

Exemplar Based WSD (k-NN)

- An exemplar based classifier is constructed for each word to be disambiguated.

  Step1: From each ***sense marked sentence*** containing the ambiguous word , a training example is constructed using:
  - POS of ***w*** as well as POS of neighboring words. - Local collocations
  - Co-occurrence vector - Morphological features
  - Subject-verb syntactic dependencies

  Step2: Given a test sentence containing the ambiguous word, a test example is similarly constructed.

  Step3: The test example is then compared to all training examples and the k-closest training examples are selected.

  Step4: The sense which is most prevalent amongst these "k" examples is then selected as the correct sense.

Beyond Supervised Methods

- If you want to be able to do WSD in the large, you need to be able to disambiguate all words in a text
- It is hard to get large amount of annotated data for every word in text
  - Use existing manually tagged data (SENSEVAL-2, 5000 words from Penn Treebank)
  - Use parallel bilingual data
  - Check OpenMind Word Expert project

http://teach-computers.org/word-expert.html

Semi-supervised Decision List Algorithm

- Based on Yarowsky's supervised algorithm that uses *Decision Lists.*
- Collecting seed examples
  - Goal: start with a small subset of the training data being labeled
  - Label a number of training examples by hand
  - Pick a single feature for each class by hand
  - Use words in dictionary definitions
    - a vegetable organism, ready for planting
    - equipment, machinery, apparatus, for industrial plant

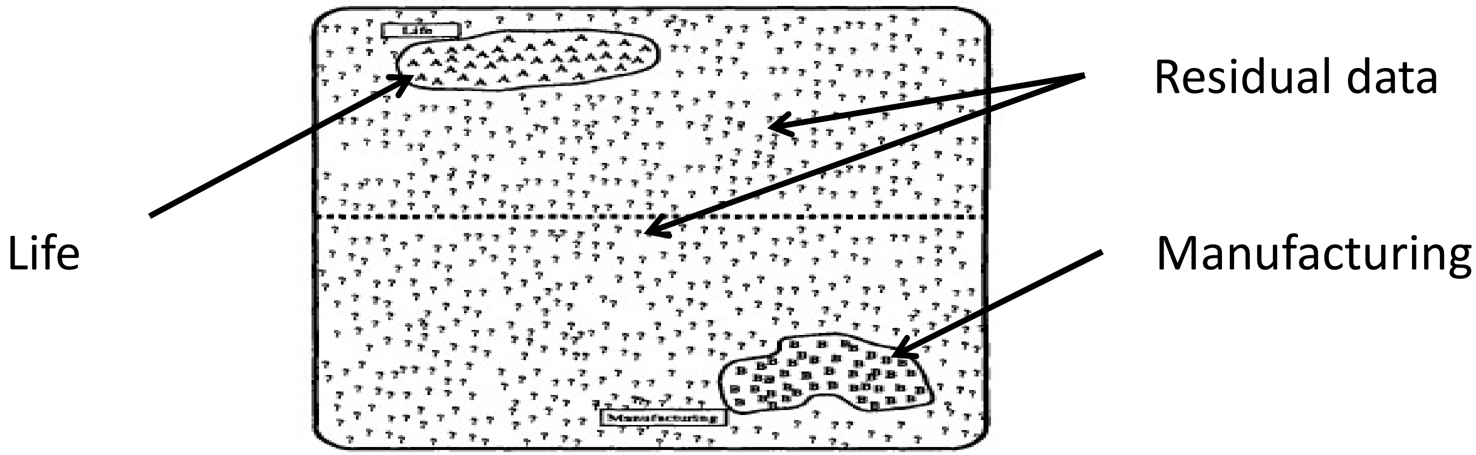
Collecting Seed Examples:

- For the “plant” sense distinction, initial seeds are“word-within-k=life” and“word-within-k=manufacturing”
- Partition the unlabeled data into three sets:

- 82 examples labeled with “life” sense
- 106 examples labeled with “manufacturing” sense
- 7350 unlabeled examples

Iterative Bootstrapping – Step 1

- Identify all contexts in which the polysemous word occurs.
- For each possible sense use seed collocations to identify a relatively small number of training examples representative of that sense.

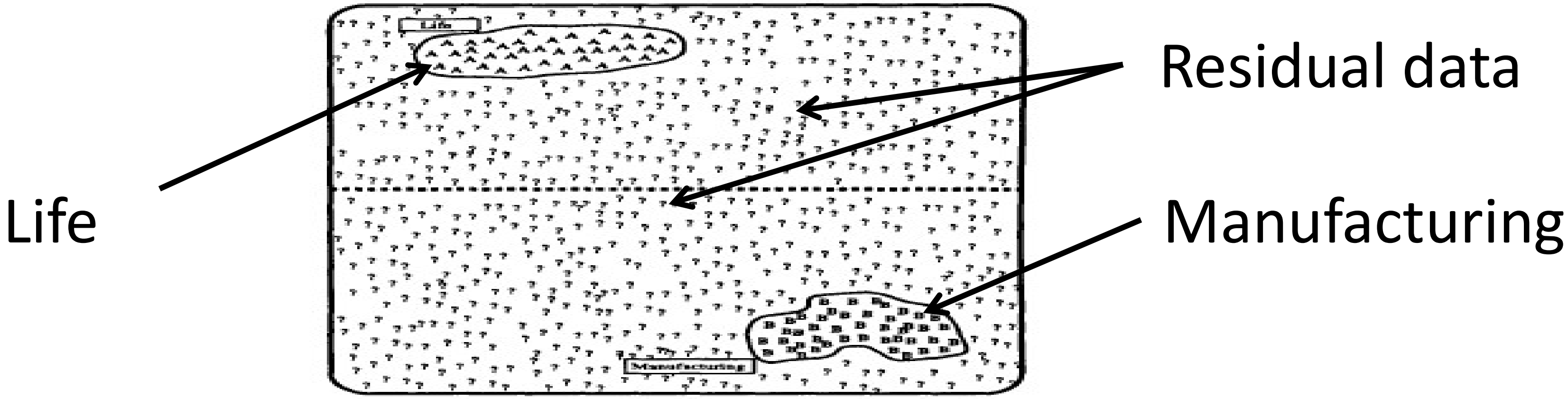


- Seed collocation should accurately distinguish the senses.

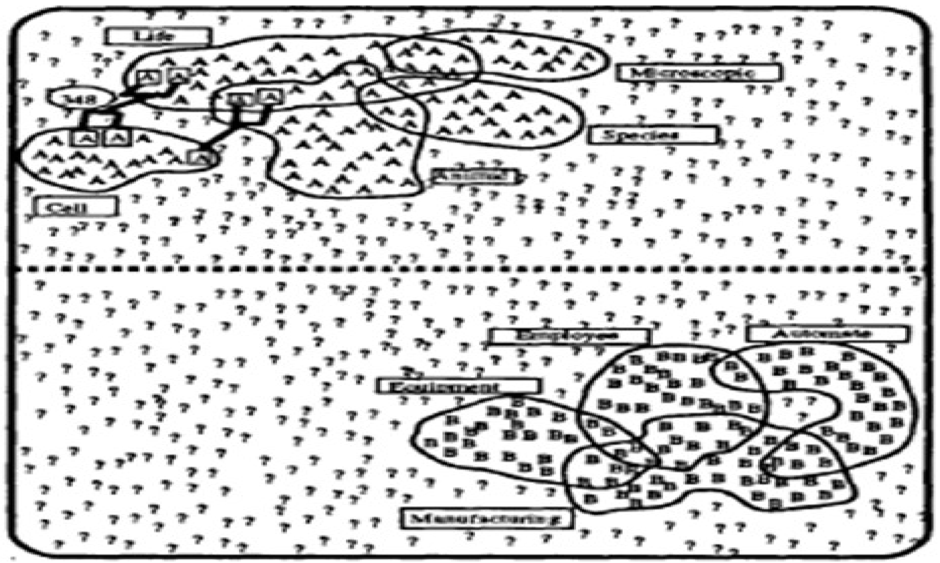
Iterative Bootstrapping – Step 2

- Train the *Decision List* algorithm on the seed data.
- Classify the entire sample set using the trained classifier.
- Create new seed data by adding those members which are tagged as Sense-A or Sense-B with high probability.
- Retrain the classifier using the new seed data.
- These additions will contribute new collocations that are reliably indicative of the 2 senses.

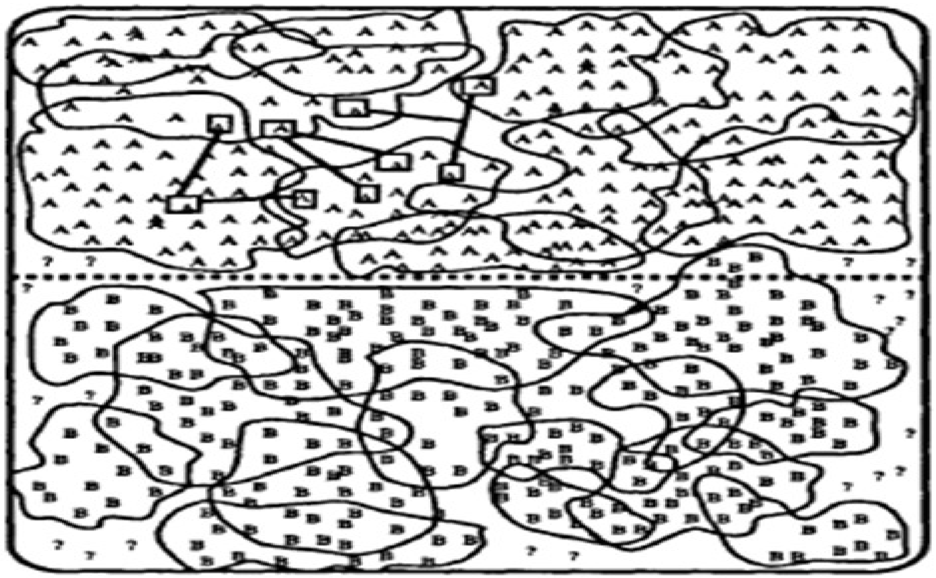
Initialization, Progress and Convergence



Seed set grows



Stop when residual set



WSD: Unsupervised Approach Using ROGET'S Thesaurus Categories

- Based on three observations:
  - Different conceptual classes of words tend to appear in recognizably different contexts.
  - Different word senses belong to different conceptual classes (E.g. crane 吊车/ 鹤)
  - A context based discriminator for the conceptual classes can serve as a context based discriminator for the members of those classes.
- Identify ***salient*** words in the collective context of the thesaurus category and weigh appropriately.

$$\text{Weight(word)} = \text{Salience(Word)} = \frac{\Pr(w|RCat)}{\Pr(w)}$$

ANIMAL/INSECT
species (2.3), family(1.7), bird(2.6), fish(2.4), egg(2.2), coat(2.5), female(2.0), eat (2.2), nest(2.5), wild

TOOLS/MACHINERY
tool (3.1), machine(2.7), engine(2.6), blade(3.8), cut(2.2), saw(2.5), lever(2.0), wheel (2.2), piston(2.5)

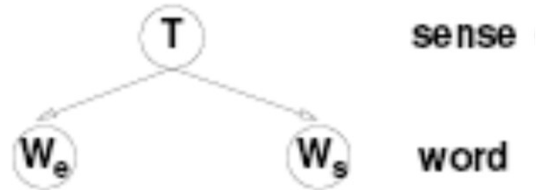
- Predict the appropriate category for an ambiguous word using the weights of words in its context: argmax $\sum_{w \text{ in context}} \log\left(\frac{\Pr(w|Rcat) * \Pr(Rcat)}{\Pr(w)}\right)$

…lift water and to grind grain. Treadmills attached to *cranes* were used to lift heavy objects from Roman times, ….

TOOLS/MACHINE	Weight	ANIMAL/INSECT	Weight
lift	2.44	Water	0.76
grain	1.68		
used	1.32		
heavy	1.28		
Treadmills	1.16		
attached	0.58		
grind	0.29		
Water	0.11		
TOTAL	**11.30**	**TOTAL**	**0.76**

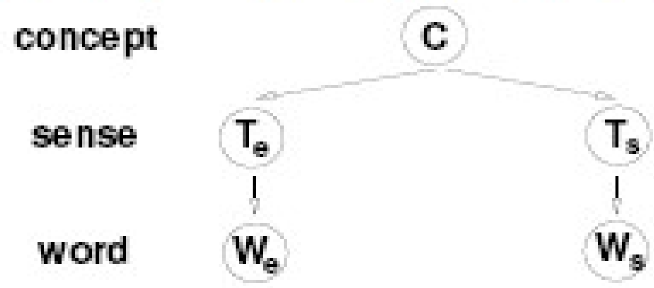
WSD: Using Bilingual Corpora

- A word having multiple senses in one language will have distinct translations
- The translations are considered as contextual indicators of the sense
- **<u>Sense Model</u>**



$$P(W_e, W_s, T) = P(T).P(W_e|T).P(W_s|T) \quad \textbf{(5.1)}$$

- **<u>Concept Model</u>**



$$P\left(W_e, W_s, T_s, T_e, C\right) = P(C).P(T_e|C).P(T_s|C).P(W_e|T_e).P(W_s|T_s) \quad \textbf{(5.2)}$$

WSD: Hybrid Approach
An Iterative Approach

- Uses semantic relations (synonymy and hypernymy) form WordNet.
- Extracts collocational and contextual information form WordNet (gloss) and a small tagged data.
- Monosemic words in the context serve as a seed set of disambiguated words.
- In each iteration new words are disambiguated based on their semantic distance from already disambiguated words.

- Precision 92.2% Recall 55%

SenseLearner

Ref

Semantic Generalizations

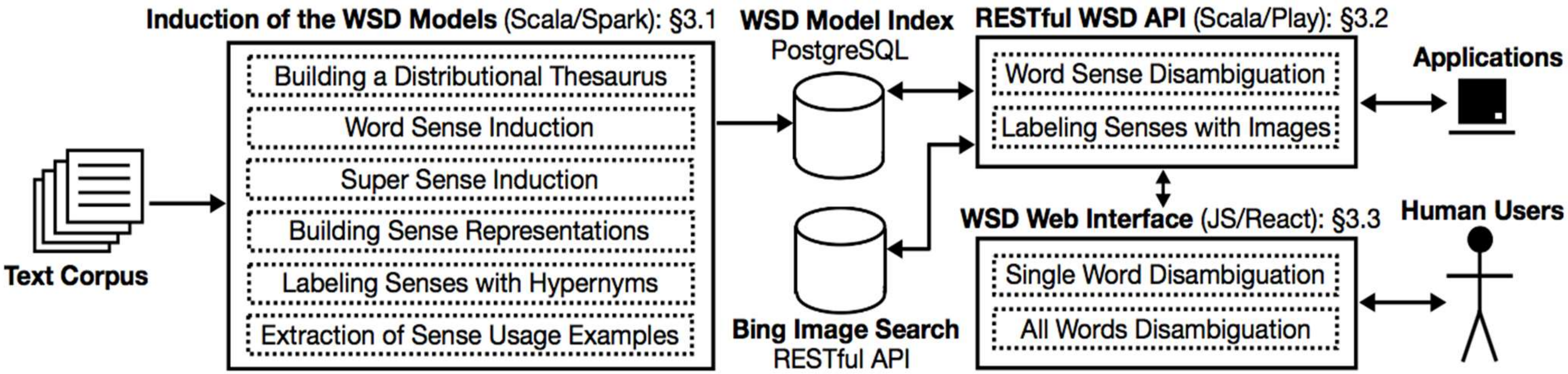
- Improvises Lin's algorithm by using semantic dependencies from WordNet.

E.g.

- if "**drink** **water**" is observed in the corpus then using the hypernymy tree we can derive the syntactic dependency "**take-in** **liquid**"
- "**take-in** **liquid**" can then be used to disambiguate an instance of the word tea as in "**take** **tea**", by using the hypernymy-hyponymy relations.
- Precision : 64.6% Recall 64.6%

Knowledge-Free and Interpretable Word Sense Disambiguation

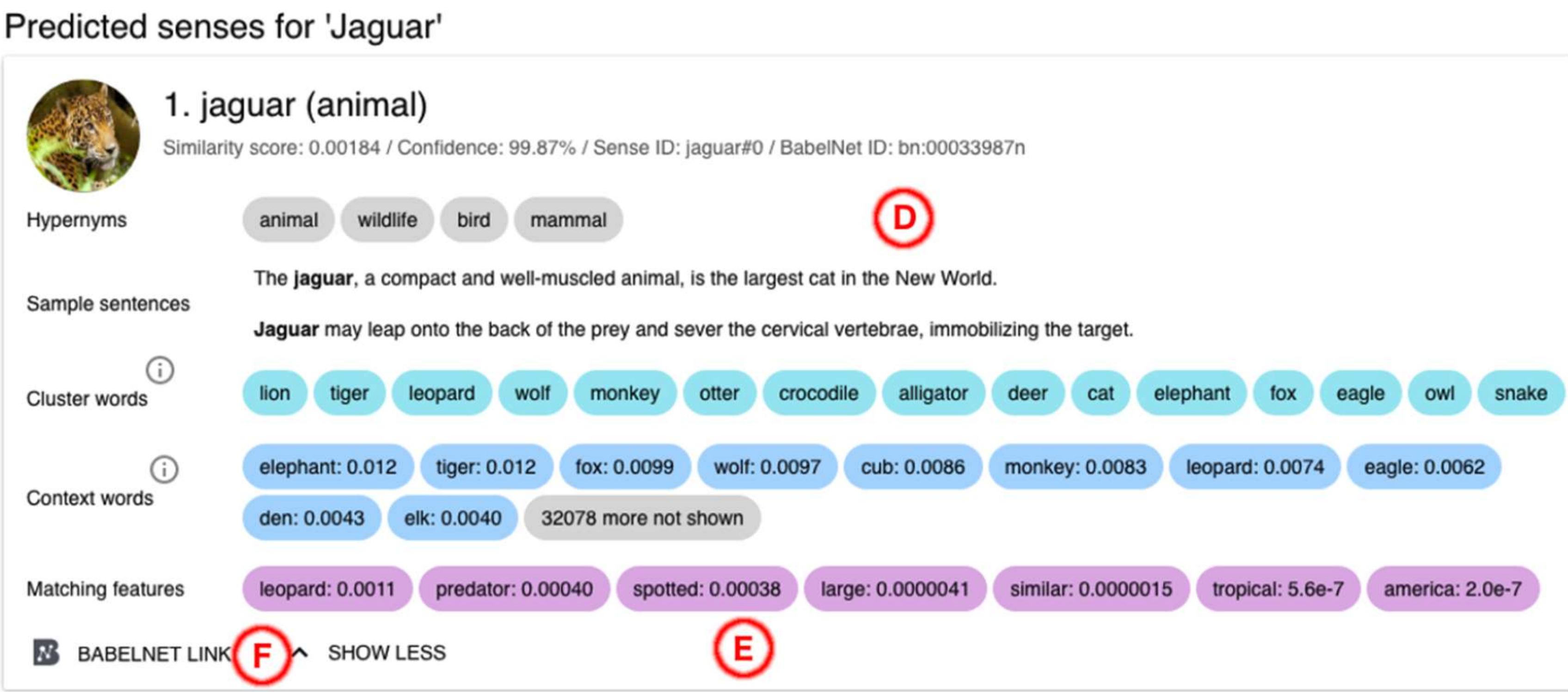
- Word senses based on cluster word features
- Word senses based on context word features
- Super senses based on cluster word features
- Super senses based on context word features



Alexander Panchenko, Fide Marten, Eugen Ruppert, Stefano Faralli, Dmitry Ustalov, Simone Paolo Ponzetto, Chris Biemann. Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 91-96.

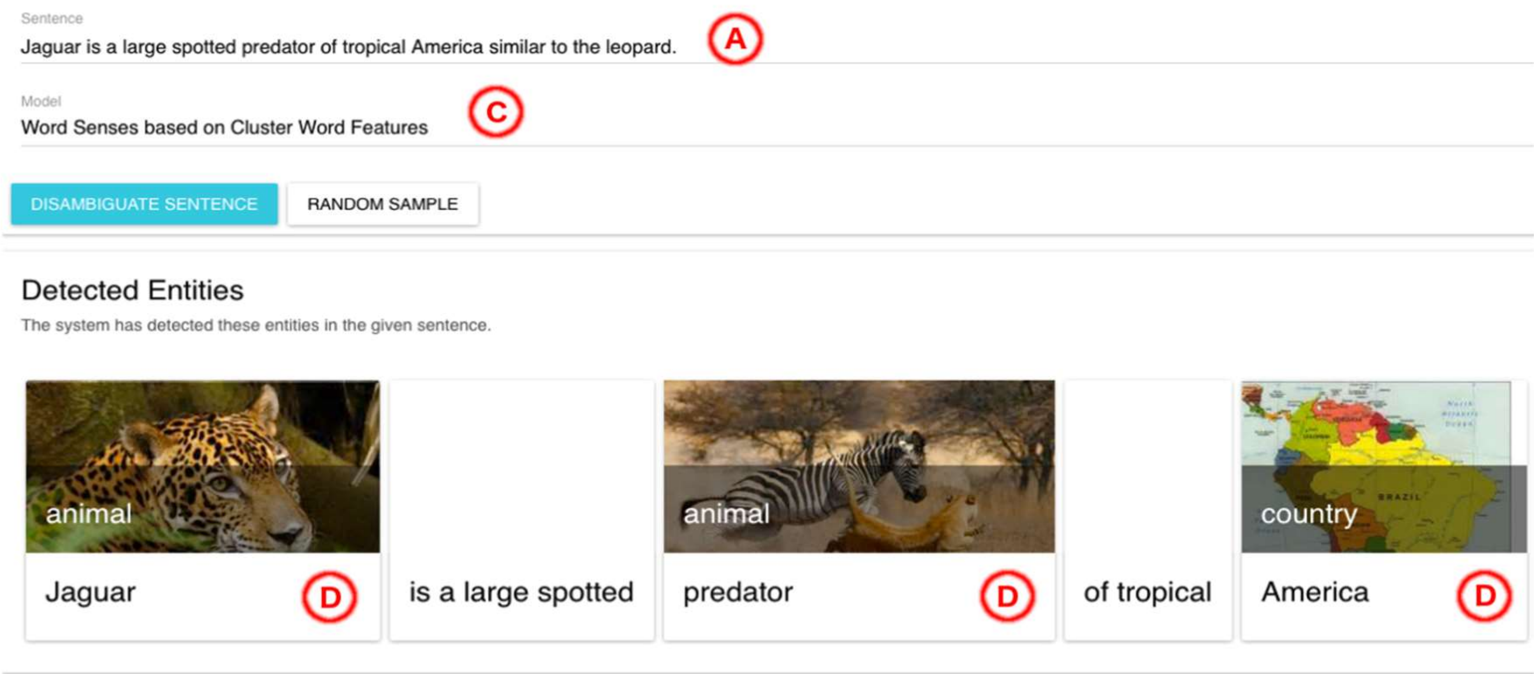
User Interface for Interpretable WSD

- **<u>Single word disambiguation mode</u>** : a user specifies an ambiguous word and its context



User Interface for Interpretable WSD

- **<u>All words disambiguation mode</u>** : the system performs disambiguation of all nouns and entities in the input text

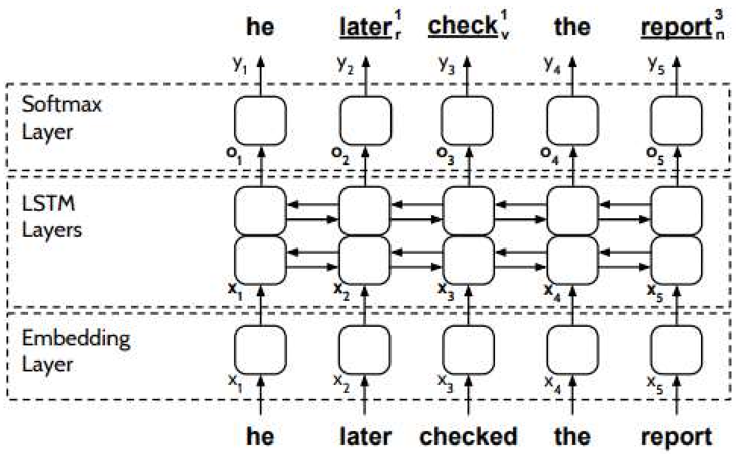


- How well hypernyms of ambiguous words are assigned in context by the system.
  E.g.
  "**animal**" for the word "**Jaguar**" in the context
  "***Jaguar*** is a large spotted predator of tropical **America**"

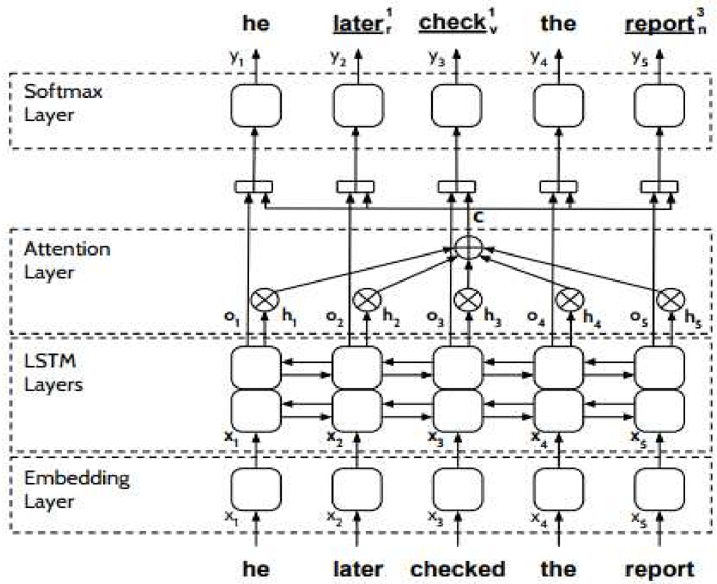
- Performance: Hypers 0.308, HyperHypers 0.686

Neural Approach

- Raganato A et al. (2017)
  - Propose and study in depth a series of end-to-end neural architectures directly tailored to the task, from bidirectional Long Short-Term Memory to encoder-decoder models.



Bidirectional LSTM sequence labeling architecture for WSD (2 hidden layers).



Attentive bidirectional LSTM sequence labeling architecture for WSD (2 hidden layers).

Raganato A, Bovi C D, Navigli R. Neural sequence learning models for word sense disambiguation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1167-1178.

Neural Approach

- Raganato A et al. (2017)

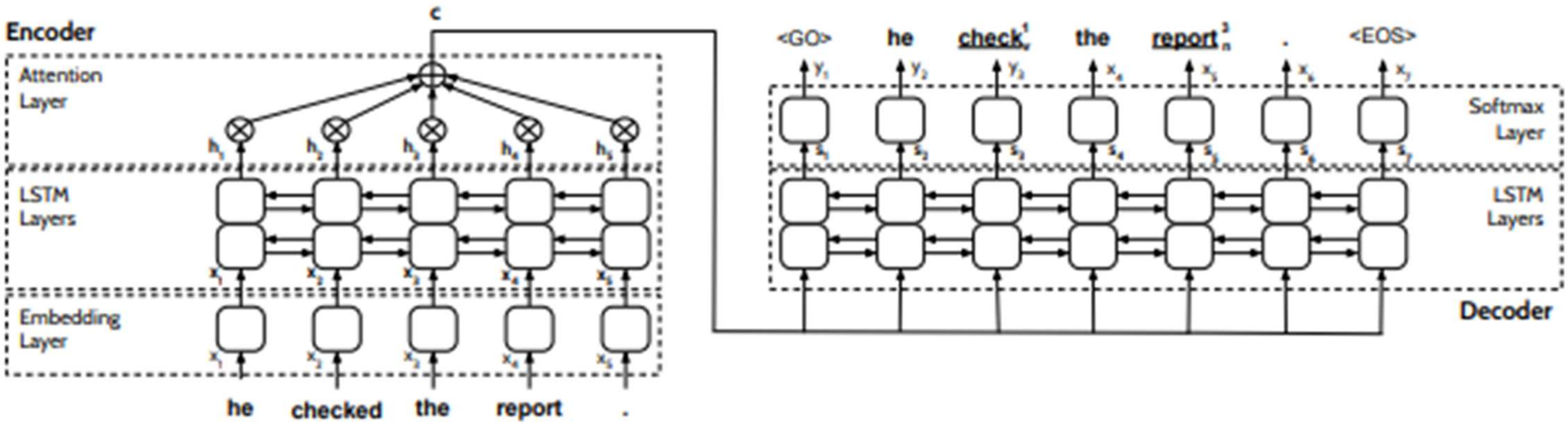
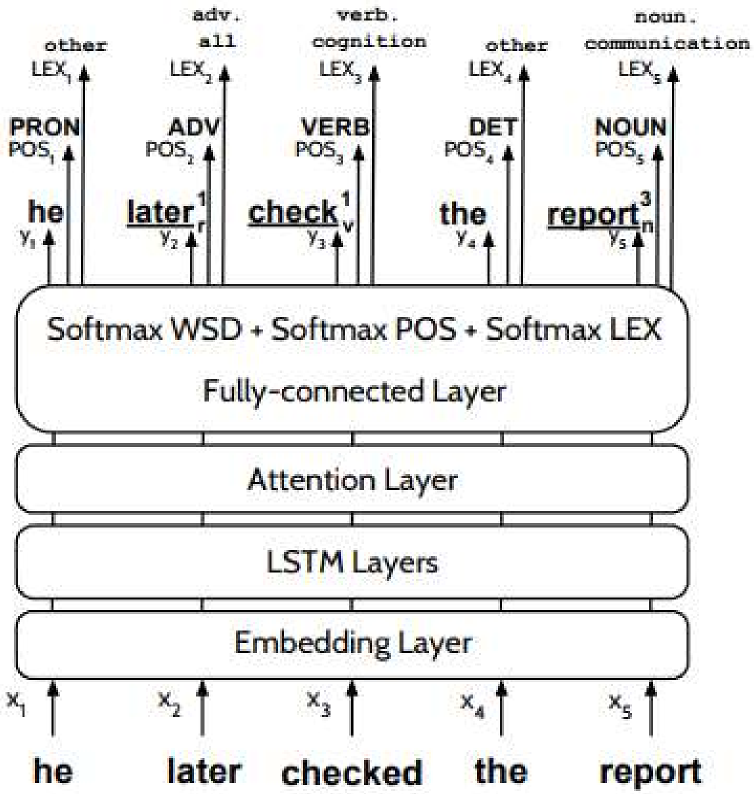


Figure 3: Encoder-decoder architecture for sequence-to-sequence WSD, with 2 bidirectional LSTM layers and an attention layer.

Raganato A, Bovi C D, Navigli R. Neural sequence learning models for word sense disambiguation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1167-1178.

Neural Approach

- Raganato A et al. (2017)

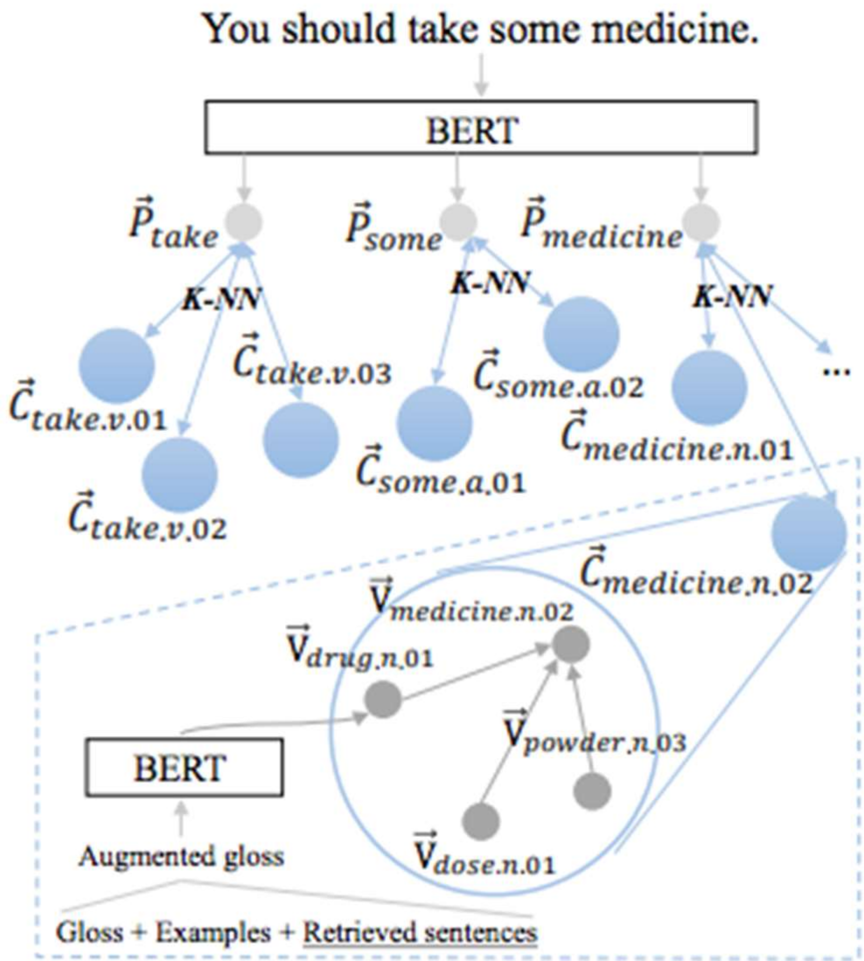


Multitask augmentation (with both POS and LEX as auxiliary tasks) for the attentive bidirectional LSTM tagger

Raganato A, Bovi C D, Navigli R. Neural sequence learning models for word sense disambiguation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1167-1178.

Neural Approach

- Contextual embeddings
  - [Wang et al. 2020] propose a Synset Relation-Enhanced Framework (SREF) that leverages sense relations for both sense embedding enhancement and a try-again mechanism that implements WSD again, after obtaining basic sense embeddings from augmented WordNet glosses.



Wang, M. and Wang, Y., 2020, November. A Synset Relation-enhanced Framework with a Try-again Mechanism for Word Sense Disambiguation.

Overcoming Knowledge Bottle-Neck

- Using Search Engines
  - Construct search queries using monosemic words and phrases form the gloss of a synset.
  - Feed these queries to a search engine.
  - From the retrieved documents extract the sentences which contain the s earch queries
- Using Equivalent Pseudo Words
  - Use monosemic words belonging to each sense of an ambiguous word.
  - Use the occurrences of these words in the corpus as training examples for the ambiguous word

Bounds

- Measure of how well the algorithm performs relative to the difficulty of the task.
- ***Upper Bound***: Human performance.
- ***Lower Bound or baseline***: Usually the assignment of all contexts to the most frequent sense.

Senseval Competition

- Comparison of various systems, trained and tested on the same set
- Senses are selected from WordNet
- Sense-tagged corpora available
  - http://www.itri.brighton.ac.uk/events/senseval

- SensEval I (1998)
  - Overall: 75%   Nouns: 80% Verbs: 70%
- SemEval series

The Next Lecture

- Lecture 9
  Language Model