# WSD: Unsupervised

Unsupervised

It is expensive and difficult to build large corpora in which each word is labeled for its word sense.

In unsupervised approaches, we start with no known answers, and no predefined senses.

The set of "senses" is created automatically from the instances of each word in the training set.

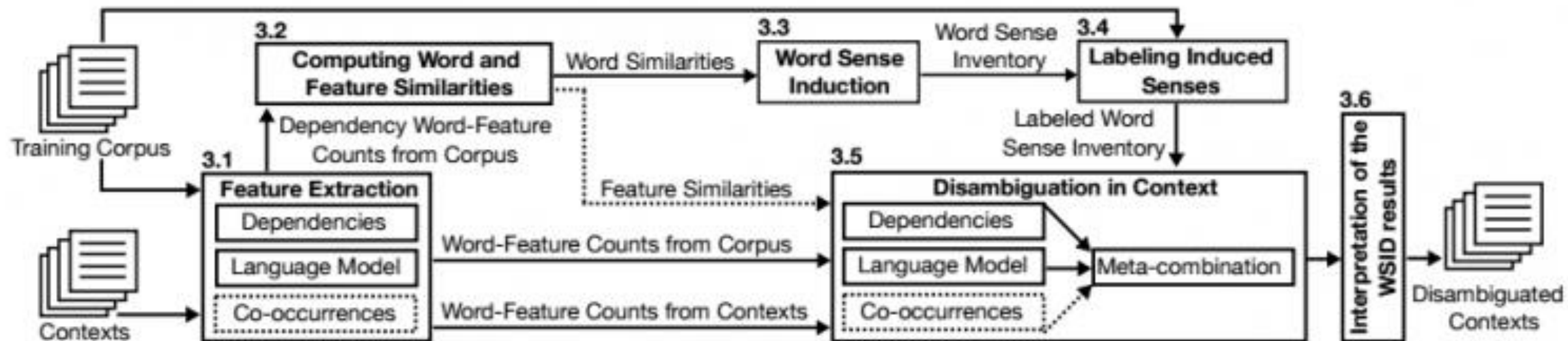# Word sense induction: **clustering**

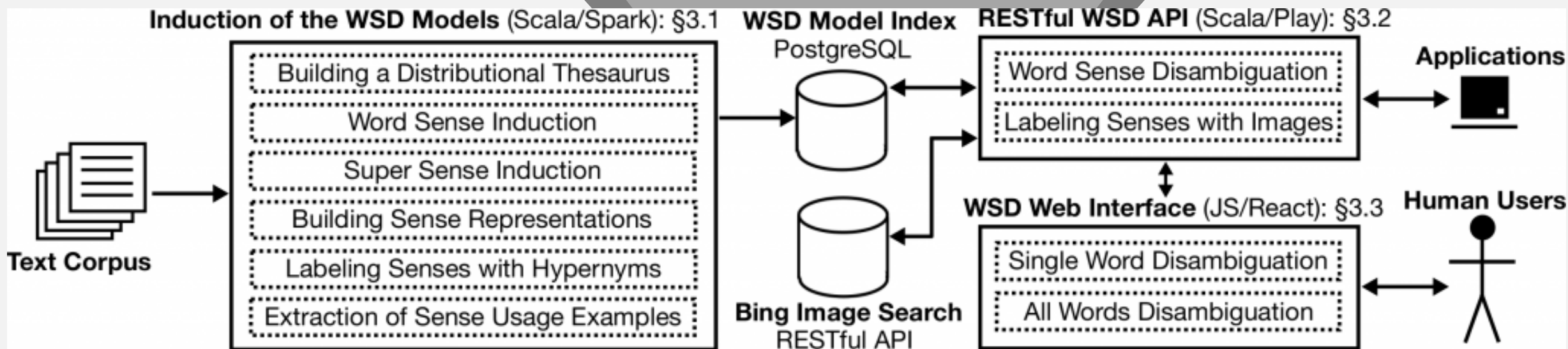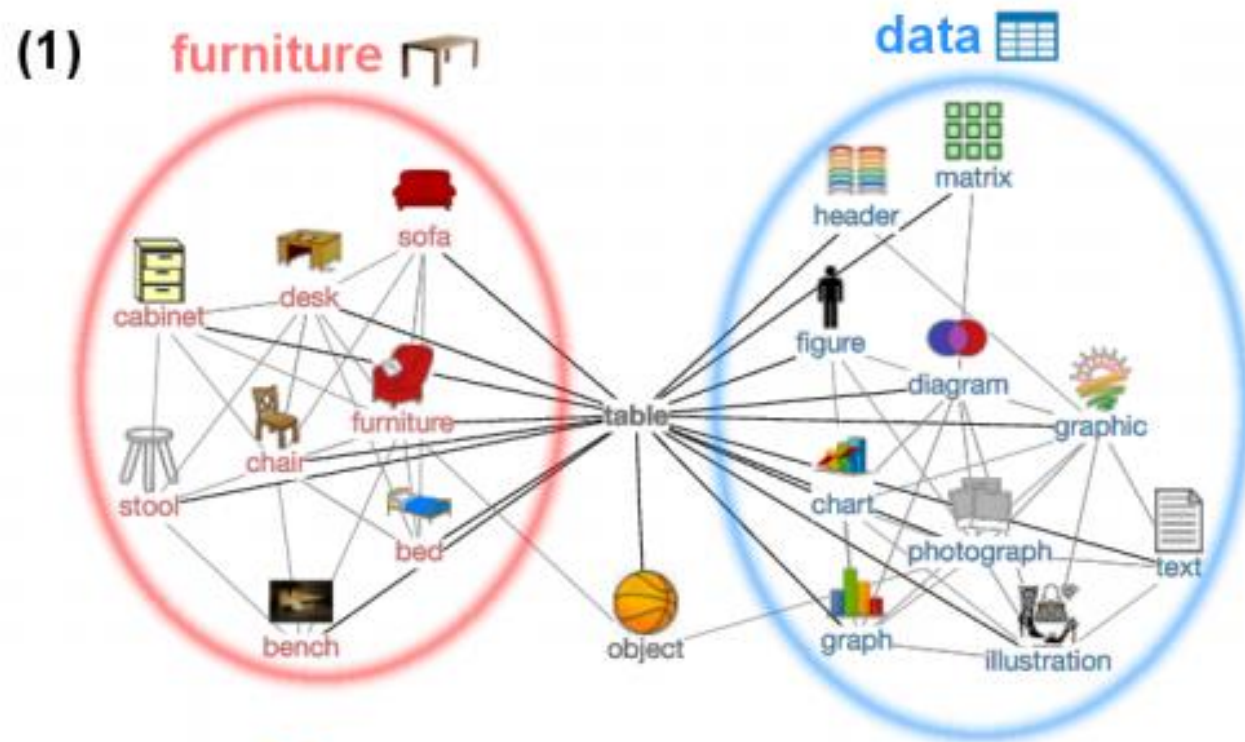# Word sense induction: clustering Disambiguate
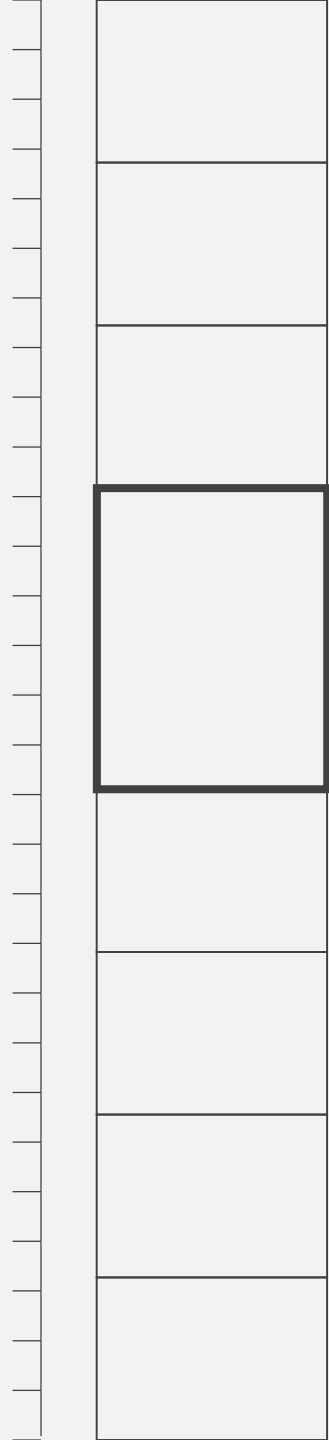
JoBim Text

Method

1. extraction of context features
2. computing word and feature similarities
3. word sense induction
4. labeling of clusters with hypernyms and images
5. disambiguation of words in context
   based on the induced inventory
6. Interpretation of the model

Induction of the WSD Models (Scala/Spark): §3.1

- Building a Distributional Thesaurus
- Word Sense Induction
- Super Sense Induction
- Building Sense Representations
- Labeling Senses with Hypernyms
- Extraction of Sense Usage Examples

Text Corpus

WSD Model Index
PostgreSQL

Bing Image Search
RESTful API

RESTful WSD API (Scala/Play): §3.2

- Word Sense Disambiguation
- Labeling Senses with Images

Applications

WSD Web Interface (JS/React): §3.3

- Single Word Disambiguation
- All Words Disambiguation

Human Users

Extraction of context features

1. Stanford Dependencies
2. Co-occurrence features
3. Tri-gram features

(1) furniture 🪑   data 📊

furniture group: cabinet, sofa, desk, furniture, chair, stool, bed, bench

table

data group: header, matrix, figure, diagram, graphic, chart, photograph, graph, illustration, text, object

(2)

furniture 🪑	data 📊
seating#NN#conj_and	graph#NN#-conj_and
counter#NN#-conj_and	index#NN#-conj_and
bench#NN#-conj_and	knot#NN#nn
folding#JJ#amod	row#NN#prep_with
desk#NN#-conj_and	graph#NN#appos
lunch#NN#nn	above#JJ#dep
furnish#VB#-prep_with	above#RB#-npadvmod
sofa#NN#-conj_and	diagram#NN#conj_and
booth#NN#-conj_and	below#IN#amod
stool#NN#conj_and	draw#VB#amod
armchair#NN#-conj_and	reproduce#VB#rcmod
...	...

(3)

📊 —data | dep—

data | LM   data | cooc   furniture | cooc   data | cooc   data | cooc

Below is a table of   information , which lists all   cookies   deployed and used on our   website   .

Algorithm 1: Unsupervised WSD of the word t based on the induced word sense inventory I.

---

input : Word t, context features C, sense inventory I, word-feature table F, use largest cluster back-off LCB, use feature expansion FE.

output: Sense of the target word t in inventory I and confidence score.

1   $S \leftarrow \mathrm{getSenses}\,(I, t)$
2   **if** FE **then**
3 | $C \leftarrow \mathrm{featureExpansion}(C)$
4   **end**
5   **foreach** $(sense, cluster) \in S$ **do**
6 | $\alpha[sense] \leftarrow \{\}$
7 | **foreach** $w \in cluster$ **do**
8 | | **foreach** $c \in C$ **do**
9 | | | $\alpha[sense] \leftarrow \alpha[sense] \cup F(w, c)$
10 | | **end**
11 | **end**
12 **end**
13 **if** $\max_{sense \in S}\, mean(\alpha[sense]) = 0$ **then**
14 | **if** LCB **then**
15 | | **return** $\arg\max_{(\_,cluster) \in S} |cluster|$
16 | **else**
17 | | **return** -1 // reject to classify
18 | **end**
19 **else**
20 | **return** $\arg\max_{(sense,\_) \in S} mean(\alpha[sense])$
21 **end**

The algorithm starts by retrieving induced sense clusters of the target word (line 1). Next, the method starts to accumulate context feature weights of each sense in α[sense]. Each word w in a sense cluster brings all its word-feature counts F(w, c): see lines 5-12. Finally, a sense that maximizes mean weight across all context features is chosen (lines 13-21). Optionally, we can resort to the largest cluster back-off (LCB) strategy in case if no context features match sense representations.

http://ltbev.informatik.uni-hamburg.de/wsd/single-word

Try

Table:

He spread out the newspaper on the table.
The offer on the table is a 10% wage increase.
Apply the style to each table in the document.

Unsupervised Knowledge Free Word Sense Disambiguation

1. table (item)

Similarity score: 0.000131 / Confidence: 99.07% / Sense ID: table#1 / BabelNet ID: bn:00075813n

He spread out the newspaper on the table.

Hypernyms
item   furniture   product   object

Related words
aka Cluster words

desk   the table   kitchen table   chair   dining table   coffee table   sofa   bed   shelf   the counter   furniture

the desk   couch   stool   dinner table   tray   seating   floor   fireplace   dresser   130 more not shown

Features
aka Context features

live: 0.000049   say: 0.00013   open: 0.00097   go: 0.00077   sit: 0.018   kill: 0.000027   hold: 0.00023

close: 0.000018   pay: 0.00014   send: 0.000075   see: 0.00023   offer: 0.00016   take: 0.00046   set: 0.00040

meet: 0.00026   head: 0.00039   run: 0.00042   percent: 0.0000070   make: 0.00044   fall: 0.0020

33485 more not shown

Matching features
spread: 0.000083   newspaper: 0.000047

1. table (item)

Similarity score: 0.0000416 / Confidence: 39.87% / Sense ID: table#1 / BabelNet ID: bn:00075813n

The offer on the table is a 10% wage increase.

Hypernyms
item   furniture   product   object

Related words
aka Cluster words
desk   the table   kitchen table   chair   dining table   coffee table   sofa   bed   shelf   the counter   furni

the desk   couch   stool   dinner table   tray   seating   floor   fireplace   dresser   130 more not shown

Features
aka Context features
live: 0.000049   say: 0.00013   open: 0.00097   go: 0.00077   sit: 0.018   kill: 0.000027   hold: 0.00023

close: 0.000018   pay: 0.00014   send: 0.000075   see: 0.00023   offer: 0.00016   take: 0.00046   set: 0.00040

meet: 0.00026   head: 0.00039   run: 0.00042   percent: 0.0000070   make: 0.00044   fall: 0.0020

33485 more not shown

Matching features
offer: 0.000031   increase: 0.0000023   wage: 0.0000011   10: 0.0000071

Unsupervised Knowledge Free Word Sense Disambiguation

1. table (item)

Similarity score: 0.000126 / Confidence: 57.24% / Sense ID: table#2 / BabelNet ID: bn:00011641n

Apply the style to each table in the document.

Hypernyms
item · object · element · page

Related words
aka Cluster words
Table · Tables · figure · section · Appendix · paragraph · Figure · clause · Paragraph · Clause · see section · see Section · Section · magnitude · Version · See Section · Annex

Features
aka Context features
live: 0.000076 · open: 0.000014 · kill: 0.0000079 · hold: 0.000019 · pay: 0.00021 · send: 0.000081 · see: 0.000043 · offer: 0.000029 · percent: 0.000094 · make: 0.000025 · play: 0.000011 · build: 0.00013 · raise: 0.000024 · face: 0.000029 · catch: 0.000052 · arrest: 0.00018 · show: 0.00045 · strike: 0.000043 · act: 0.000088 · try: 0.000026 · 8490 more not shown

Matching features
apply: 0.00012 · document: 0.0000081

BABELNET LINK

REFERENCE

J+M_C. 8

Panchenko, Alexander, et al. "Unsupervised Does Not Mean Uninterpretable: The Case for Word Sense Induction and Disambiguation." Conference of the European Chapter of the Association for Computational Linguistics 2017.
https://aclweb.org/anthology/E/E17/E17-1009.pdf

Jobim: http://ltmaggie.informatik.uni-hamburg.de/jobimtext/wsd/
http://ltbev.informatik.uni-hamburg.de/wsd/single-word

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

http://sambrody.info/Lectures/Brody_unsupervised_JHU.pdf

THANK YOU FOR WATCHING