Representation learning for words

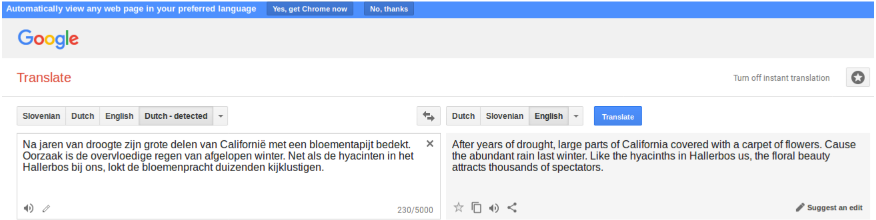
Simon Šuster
University of Antwerp & Antwerp University Hospital

`http://simonsuster.github.io/`
`@SimonSuster`

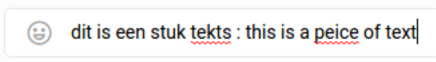
April 28, 2017

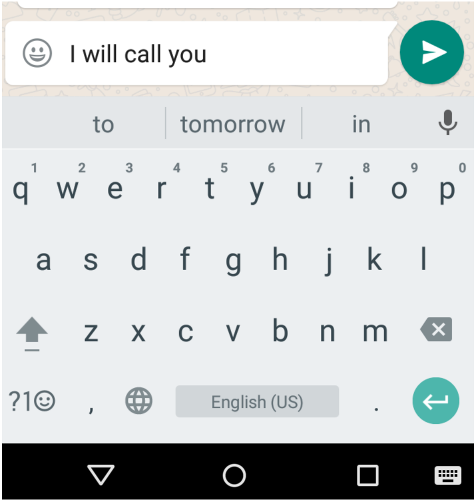
Pervasiveness of NLP

Machine translation and language detection

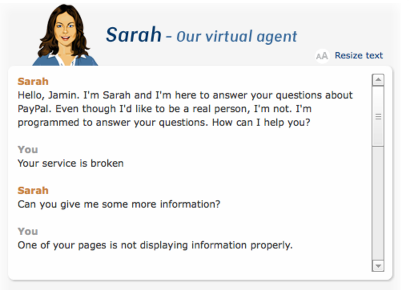


(Multilingual) spelling correction and word suggestion





QA, conversational agents and personal assistants



- Real-life applications are trained on large human-annotated datasets.
- Under the hood, low-level processing and analysis of linguistic information.

- Real-life applications are trained on large human-annotated datasets.
- Under the hood, low-level processing and analysis of linguistic information.

Most of applications work with **words** as the basic unit of text.

To a computer, text is just a long string of characters...

Necessary first steps

Pre-processing

- sentence segmentation
- tokenization
- normalization

For example:
"This is a short sentence." $\rightarrow$
["this", "be", "a", "short", "sentence", "."]

What about word meaning? How can we capture it computationally?

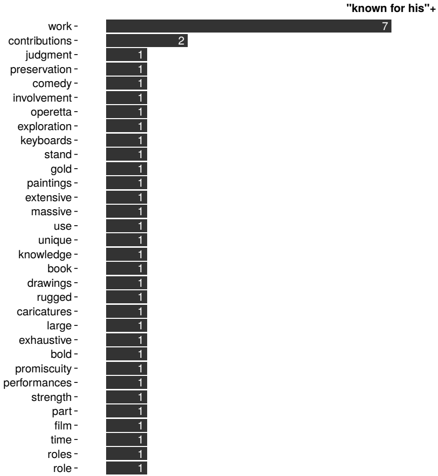
Motivating example: language models

- Estimate probabilities for all strings in a language.
- Crucial for tasks identifying words from a noisy input, in generation, in ranking word sequences.

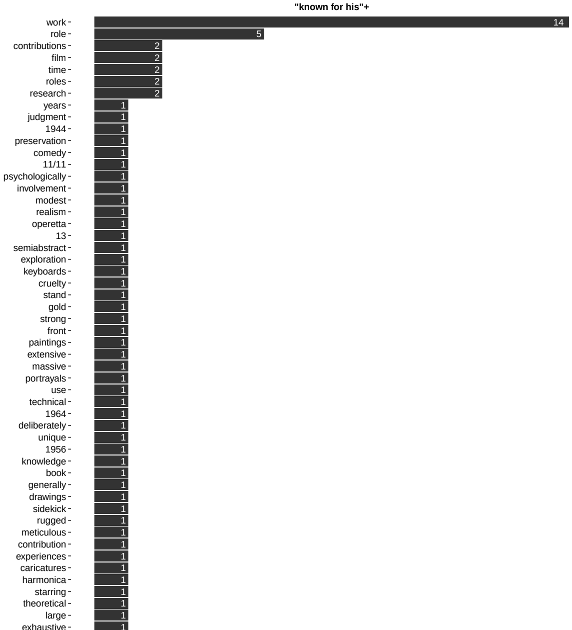
An N-gram model gives conditional probabilities:

$$p(\text{work}|\text{known}, \text{for}, \text{his}) = \frac{C(\text{known,for,his,work})}{C(\text{known,for,his})} \text{ (MLE)}$$

Estimated from 1M-word Wiki sample



Estimated from 2M-word Wiki sample



- What is $p(\text{movie}|\text{known}, \text{for}, \text{his})$ according to the above counts?

- What is $p(\text{movie}|\text{known}, \text{for}, \text{his})$ according to the above counts?
  - Answer: 0, since "known for his movie" was not observed in the data.
- Regardless of the size of the training corpus, there will always be unseen (and infrequent) words and sequences.

Lexical/data sparseness

- We need to be able to generalize and relate words
- Use the counts for "known for his film" since "movie" $\approx$ "film".

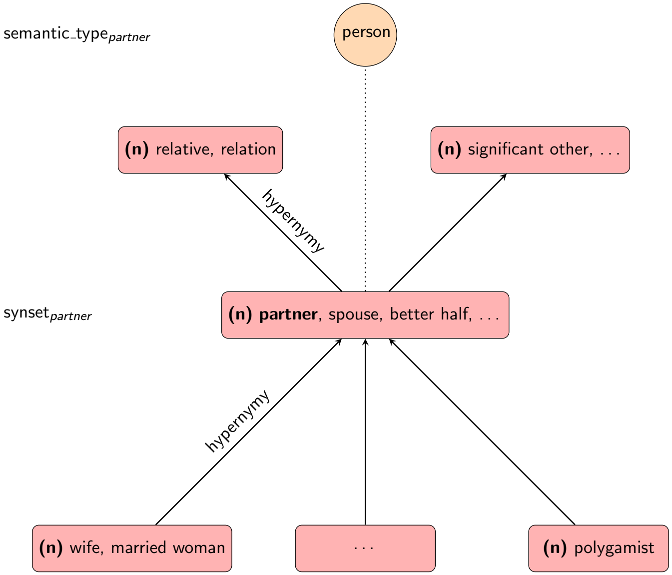
How do we obtain representations that generalize?

- Human-crafted semantic classes
- Data-induced classes and representations: **representation learning**

How do we obtain representations that generalize?

- Human-crafted semantic classes
- Data-induced classes and representations: **representation learning**
- "Specialized" representations, a mix of both (Mrkšić et al. 2017)

Human-crafted classes: WordNet



Distributional hypothesis

The meaning of a word is **an abstraction over the contexts** in which the word is used.

"You shall know a word by the company it keeps." (Firth, 1957)

Distributional hypothesis

The meaning of a word is **an abstraction over the contexts** in which the word is used.

"You shall know a word by the company it keeps." (Firth, 1957)

What's a *shrew*

An owl scooping up a **shrew**.
From where I sat, the large morsel looked remarkably like a **shrew** or baby mouse.
Underwater sniffing is not a water **shrew**'s only trick.
Shrews sometimes get into the home by falling in window wells or squeezing in tiny entry points.
What's a **shrew** and how do I get rid of them?

Small agile animal similar to a mouse

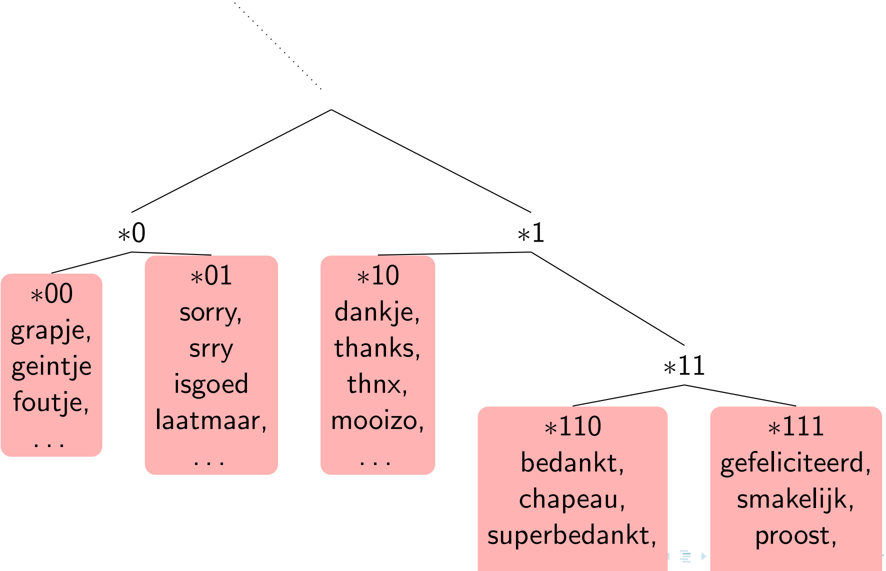
Some distributional approaches

Induce word representations from large corpora using

- clustering
- distributional semantic models (count-based)
- distributed representations (embeddings)
- latent-variable representations

Word clusters

Brown clusters from Dutch tweets



Target-context co-occurrence matrix

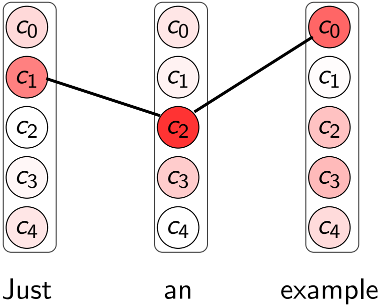
contexts

targets		leash	walk	run	owner	leg	bark
	dog	3	5	1	5	4	2
	cat	0	3	3	1	5	0
	lion	0	3	2	0	1	0
	light	0	0	0	0	0	0
	bark	1	0	0	2	1	0
	car	0	0	4	3	0	0

Word embeddings

Word	w	$C(w)$
" the "	1	[0.6762, -0.9607, 0.3626, -0.2410, 0.6636]
" a "	2	[0.6859, -0.9266, 0.3777, -0.2140, 0.6711]
" have "	3	[0.1656, -0.1530, 0.0310, -0.3321, -0.1342]
" be "	4	[0.1760, -0.1340, 0.0702, -0.2981, -0.1111]
" cat "	5	[0.5896, 0.9137, 0.0452, 0.7603, -0.6541]
" dog "	6	[0.5965, 0.9143, 0.0899, 0.7702, -0.6392]
" car "	7	[-0.0069, 0.7995, 0.6433, 0.2898, 0.6359]
...	...	...

Latent-variable representations



Learning of word representations

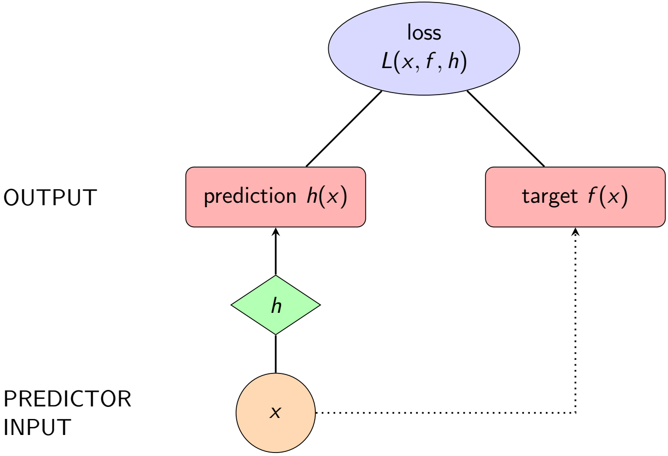
Correct solution is not knowable by humans $\rightarrow$ unsupervised learning

- Ultimately interested in extrinsic tasks
  - Features for part-of-speech tagging, named entity recognition, syntactic parsing, semantic-role labeling
- But we often measure fit to human judgments using semantic similarity benchmarks
  - A convenient and (hopefully) reliable indicator of extrinsic performance

Learning of word representations

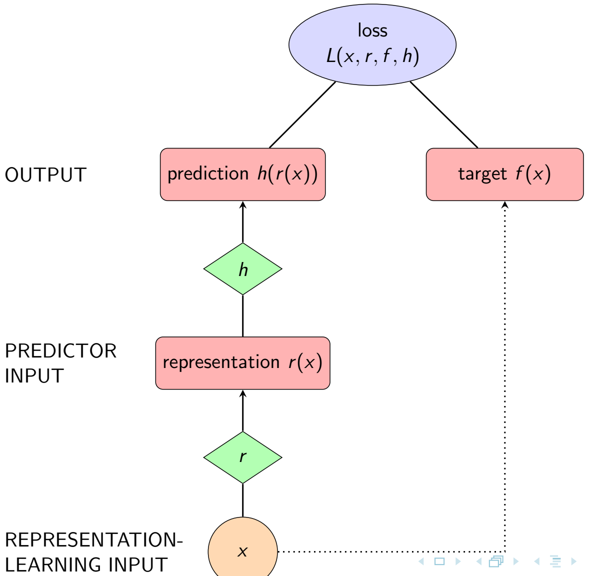
"An AI must fundamentally understand the world around us, and we argue that this can only be achieved if it can learn to identify and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data." (Bengio et al. 2013)

Supervised learning



Supervised + representation learning

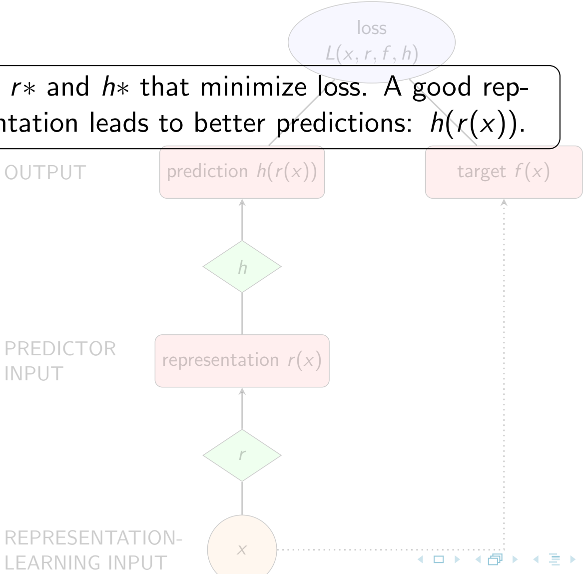
(Huang et al. 2014)



Supervised + representation learning

(Huang et al. 2014)

Find $r*$ and $h*$ that minimize loss. A good representation leads to better predictions: $h(r(x))$.



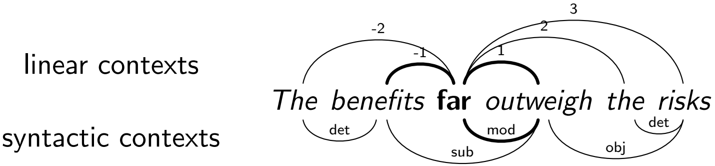
Word representations

Important areas of research

- Definition of context
- Generic vs. sense-specific
- Multilinguality

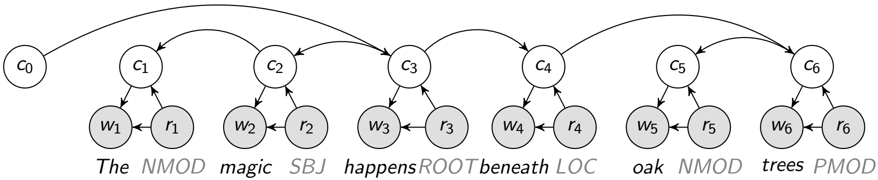
Linear vs. syntactic context

- Linear context: fixed word window to each side of the target word
- Syntactic context: follow syntactic paths ("dependencies") (Pado and Lapata 2007, Levy and Goldberg 2014)



A syntax-informed HMM model

(Šuster et al. 2015)



Sense representations

- HMMs give context-dependent representations at test time
- In other frameworks (e.g. embeddings), sense distinctions are not possible by default. Several sense-inducing extensions exist, see Camacho Collados et al. 2016 for an overview.

The jury is still out on whether sense representations are useful!

- Sense disambiguation is noisy.
- Human-defined sense distinctions not necessarily meaningful for downstream tasks.

Sense representations

Multi-sense embeddings trained on Wikipedia

rock_0	rock_1	rock_2
mud 0.897	band 0.919	disco 0.899
grass 0.877	pop 0.907	pop 0.891
deep 0.874	rapper 0.872	roll 0.883
sea 0.872	indie 0.870	gospel 0.882
cloud 0.870	punk 0.860	hip 0.867
bush 0.858	album 0.823	psychedelic 0.862
canopy 0.856	duo 0.820	hardcore 0.856
reef 0.855	supergroup 0.811	jazz 0.852
rough 0.851	singer 0.784	hop 0.847
vine 0.849	metal 0.783	contemporary 0.846
hollow 0.844	trio 0.781	mainstream 0.842
surrounding 0.841	songwriter 0.773	grunge 0.841
boulder 0.840	guitarist 0.764	techno 0.839
leaf 0.839	Pop 0.759	glam 0.837
spiral 0.839	metalcore 0.758	progressive 0.836



Multilingual representations

Goal

Obtain a representation of a concept for different languages.

- If we train (trivially) a model on different languages, the obtained parameters won't be "aligned".
- A representation for a word in the source language should be close to the representation for the word's translation in another language.
  - Requires dictionaries or word/sentence alignments.

Cross-lingual learning

Idea

Use another language to improve representations in the source language (Faruqui 2016).

Example

With multi-sense representations, we can use translations as "labels" for word senses in the source language

track: a course of study; a piece of music, a rough path...
sent_{L1}: *Choose a track that interests you*

Cross-lingual learning

Idea

Use another language to improve representations in the source language (Faruqui 2016).

Example

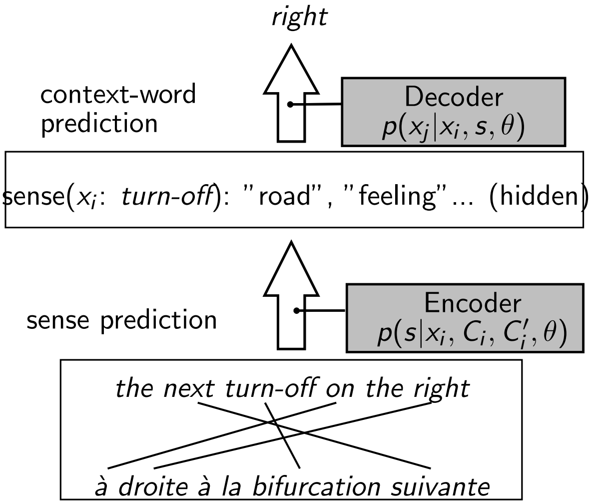
With multi-sense representations, we can use translations as "labels" for word senses in the source language

track: a course of study; a piece of music, a rough path...

sent_{L1}: *Choose a track that interests you*

sent_{L2}: *Pon una canción que te gusta*

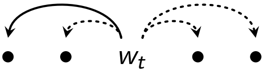
Cross-lingual learning (Šuster et al. 2016)



More on embeddings

Skip-gram embeddings (word2vec, Mikolov et al. 2013a)

- Predict context word w_c based on a target word w_t
- Consider each context separately (skip-gram)
- Input is just $< w_t, w_c >$ pairs extracted from all windows in the corpus



- Words are represented in an embedding matrix $\mathbf{W} \in \mathbb{R}^{|V|,d}$
- Distinct target and context matrices

Skip-gram embeddings

$$p(w_c = i | w_t) = \frac{e^{\mathbf{w}_{c_i} \cdot \mathbf{w}_t}}{\sum_j e^{\mathbf{w}_{c_j} \cdot \mathbf{w}_t}}$$

- Running a logistic regression
- But update weights of both embedding matrices

Hard to optimize efficiently!

- Hierarchical softmax
- Negative sampling

Negative sampling

Intuition

- Could maximize $p(D = 1|w_t, w_c)$ under current set of weights
- Yields two-class logistic regression: $\sigma(\mathbf{w}_c \cdot \mathbf{w}_t)$
- But wouldn't lead to interesting embeddings
  - Setting all $\mathbf{w}$ to be the same would maximize all dot products and give $p = 1$
- So, incorporate pairs for which $p(D = 1|w_t, w_c)$ must be low

Negative sampling

Construct negative pairs

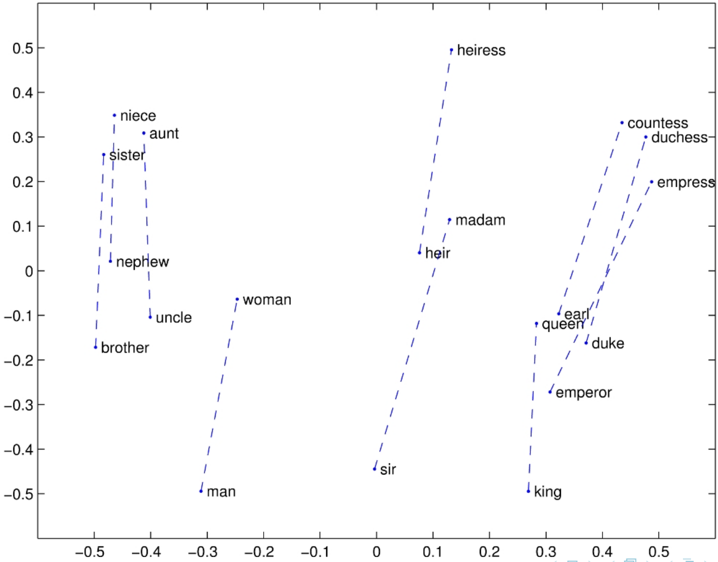
- k extra pairs per training instance
- Replacing context word with a random word

Find weights discriminating well between positive and negative pairs

- High $p(D = 1|w_t, w_c)$
- High $p(D = 0|w_t, w_{c_{rand}})$

Word analogies from embeddings

(Mikolov et al. 2013b)



Summary

- Pervasiveness of NLP
- Words as basic units
- Lexical sparseness (based on a language-model example)
- Types of word representations
- Representation learning (with its relationship to supervised learning)
- Active research areas for word representations
- Word embeddings

References

Nikola Mrkšić, Mohammad Taher Pilehvar and Ivan Vulić (2017). Word Vector Space Specialisation. Tutorial at EACL 2017.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). Efficient estimation of word representations in vector space. In ICLR Workshop Papers.

Mikolov, T., Yih, W. T., and Zweig, G. (2013b). Linguistic Regularities in Continuous Space Word Representations. In HLT-NAACL.

José Camacho-Collados, Ignacio Iacobacci, Roberto Navigli and Mohammad Taher Pilehvar (2016). Semantic Representations of Word Senses and Concepts. Tutorial at ACL 2016.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. Computational Linguistics, 33:161–199.

Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In ACL.

Faruqui, M. (2016). Diverse Context for Learning Word Representations (Doctoral dissertation, Carnegie Mellon University).

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1798-1828.

Stéphan Tulkens, Simon Šuster and Walter Daelemans (2016). Using Distributed Representations to Disambiguate Biomedical and Clinical Concepts. In BioNLP.

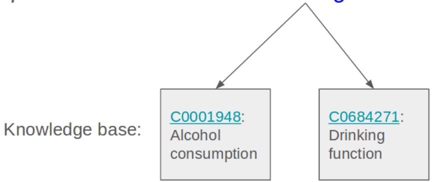
Image courtesy

http://ichef.bbci.co.uk/naturelibrary/images/ic/credit/640x395/e/el/elephant_shrew/elephant_shrew_1.jpg: 14
http://afropolitain-magazine.com/wp-content/uploads/2017/01/afropolitain-magazine-the-afro-of-today-for-tomorrow-le-poid-des-mots.jpg: 5
Stefan Evert:17
Hugo Larochelle: 18
https://nlp.stanford.edu/projects/glove/: 37

Application: Concept disambiguation (Tulkens et al. 2016)

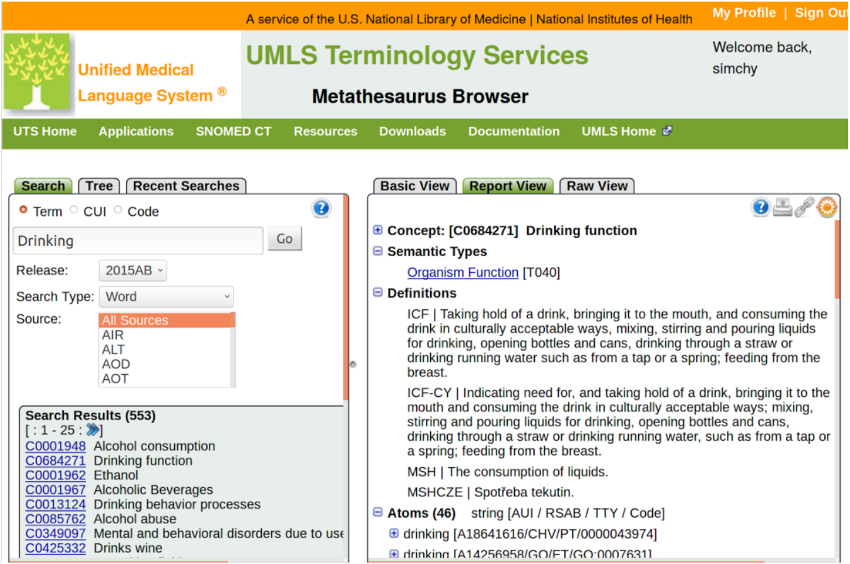
Example

"366 class 1 and 2 pupils completed a questionnaire about their drinking habits"



Idea

- Choose the sense whose KB definition is the most similar to the word's current neighborhood
- Similarly to the Simplified Lesk algorithm for word-sense disambiguation


A service of the U.S. National Library of Medicine | National Institutes of Health
My Profile | Sign Ou
Unified Medical Language System ®
UMLS Terminology Services
Metathesaurus Browser
Welcome back, simchy
UTS Home
Applications
SNOMED CT
Resources
Downloads
Documentation
UMLS Home
Search
Tree
Recent Searches
Term
CUI
Code
Drinking
Go
Release:
2015AB
Search Type:
Word
Source:
All Sources
AIR
ALT
AOD
AOT
Search Results (553)
[: 1 - 25 :]
C0001948 Alcohol consumption
C0684271 Drinking function
C0001962 Ethanol
C0001967 Alcoholic Beverages
C0013124 Drinking behavior processes
C0085762 Alcohol abuse
C0349097 Mental and behavioral disorders due to use
C0425332 Drinks wine
Basic View
Report View
Raw View
Concept: [C0684271] Drinking function
Semantic Types
Organism Function [T040]
Definitions
ICF | Taking hold of a drink, bringing it to the mouth, and consuming the drink in culturally acceptable ways, mixing, stirring and pouring liquids for drinking, opening bottles and cans, drinking through a straw or drinking running water such as from a tap or a spring; feeding from the breast.
ICF-CY | Indicating need for, and taking hold of a drink, bringing it to the mouth and consuming the drink in culturally acceptable ways; mixing, stirring and pouring liquids for drinking, opening bottles and cans, drinking through a straw or drinking running water, such as from a tap or a spring; feeding from the breast.
MSH | The consumption of liquids.
MSHCZE | Spotřeba tekutin.
Atoms (46) string [AUI / RSAB / TTY / Code]
drinking [A18641616/CHV/PT/0000043974]
drinking [A14256958/GO/ET/GO:0007631]

Procedure

1. Train biomedical embeddings
2. Based on the embeddings and the UMLS thesaurus, represent each concept s with a vector v_s:
   - v_s: is the average of definition vectors d_s
   - d_s: is the sum over vectors of all words in the definition
3. For every occurrence of an ambiguous word w in a document, sum the vectors of context words
4. Average these summed vectors into x_w
5. Choose the highest-scoring concept: $\text{argmax}_s \text{cosine}(v_s, x_w)$