CS6120: Lecture 10 Lexical Semantics

Kenneth Church

https://kwchurch.github.io/

Open letter to all EU leaders

GARY MARCUS

NOV 20

READ IN APP ↗

20 November 2023

Dear European leaders,

The recent events at OpenAI are likely going to lead to considerable, unpredictable instability.

The schisms on display there highlight the fact that we cannot rely purely on the companies to self-regulate AI, wherein even their own *internal* governance can be deeply conflicted.

Please don't gut the EU AI Act; we need it now more than ever.

Sincerely,

Gary Marcus

***Gary Marcus** is a leading expert on AI who testified to the US Senate Judiciary Subcommittee. An Emeritus Professor at NYU, he is the author of five books, and CEO Founder of two AI companies, one acquired by Uber.*

Please consider sharing this post.

Share

Commercial Applications: Lexical Semantics

- Ad for Ground News: https://youtu.be/nPZPrs2Uf_g?t=1137
- BOTUS: https://www.npr.org/sections/money/2017/04/07/522897876/meet-botus-planet-money-s-stock-trading-twitter-bot

Knowledge Acquisition Bottleneck: Bar-Hillel (1960) Word-Sense Disambiguation (WSD) is "AI Complete"

1. Bar-Hillel's Characterization of the Word-Sense Disambiguation Problem

Word sense disambiguation has been recognized as a major problem in natural language processing research for over forty years. One can find a number of early references, e.g., Kaplan (1950), Yngve (1955), Bar-Hillel (1960), Masterson (1967). Early on, there was a clear awareness that word-sense disambiguation is an important problem to solve: "The basic problem in machine translation is that of multiple meaning" (Masterson, 1967). But unfortunately, there was also a clear awareness that the problem is very difficult. Bar-Hillel, who had been one of the early leaders in machine translation, abandoned the field when he could not see how a program could disambiguate the word *pen* in the very simple English discourse:

Little John was looking for his toy box.
Finally he found it.
The box was in the pen.
John was very happy.

Bar-Hillel (1960, p. 159) argued that:

Assume, for simplicity's sake, that *pen* in English has only the following two meanings: (1) a certain writing utensil, (2) an enclosure where small children can play. I now claim that no existing or imaginable program will enable an electronic computer to determine that the word *pen* in the given sentence within the given context has the second of the above meanings, whereas every reader with a sufficient knowledge of English will do this "automatically."

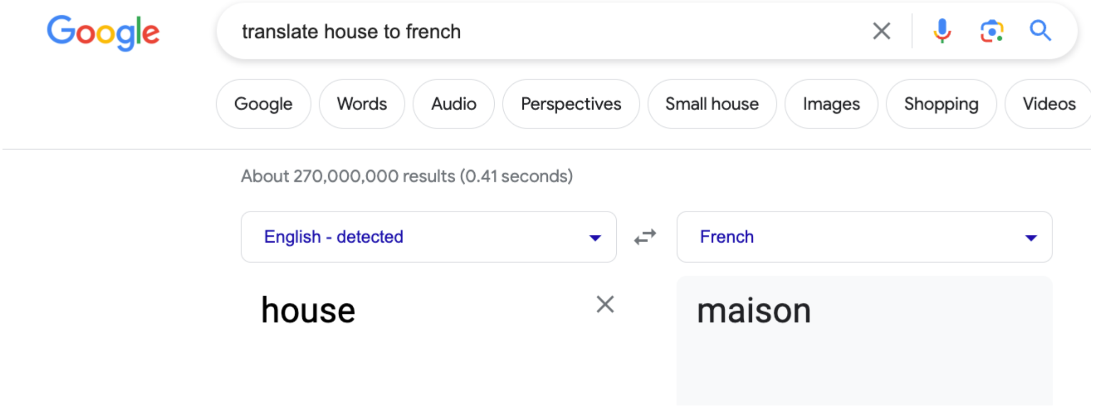
A method for disambiguating word senses in a large corpus (Gale, Church & Yarowsky, 1991)

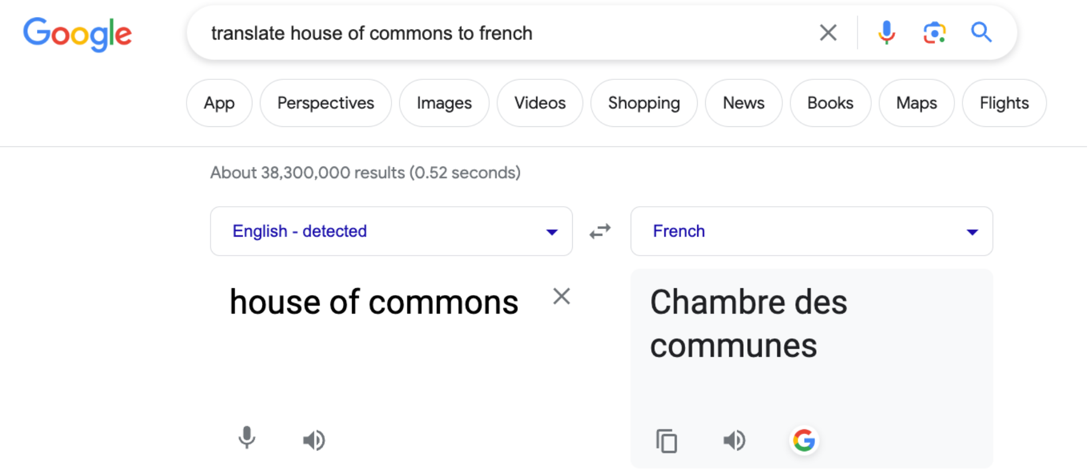
Paths Forward

- Tasks:
  - Word-Sense Disambiguation, Metaphor
  - Idioms
  - NER (Named Entity Recognition)
  - Linking
  - …

- Rules
  - Assume productive processes (e.g., compositionality)
- Lexical Resources
  - Dictionaries,
  - Ontologies (WordNet, Cyc)
- Corpora
- Large Language Models (LLMs)

house → maison | Chambre





We took the initiative in assessing and amending current
pris initiative evaluer modifier

legislation and policies to ensure that they reflect
lois politiques afin correspondent

a broad interpretation of the charter
genereuse interpretation charte

Table IV: A Contingency Table

	chambre	
house	31,950	12,004
	4,793	848,330

Table I: Contextual Clues for Sense Disambiguation		
Word	Sense	Contextual Clues
drug	medicaments	prices, prescription, patent, increase, generic, companies, upon, consumers, higher, price, consumer, multinational, pharmaceutical, costs
drug	drogues	abuse, paraphernalia, illicit, use, trafficking, problem, food, sale, alcohol, shops, crime, cocaine, epidemic, national, narcotic, strategy, head, control, marijuana, welfare, illegal, traffickers, controlled, fight, dogs
sentence	peine	inmate, parole, serving, a, released, prison, mandatory, judge, after, years, who, death, his, murder
sentence	phrase	I, read, second, amended, ", ", protects, version, just, letter, quote, word, ..., last, amendment, insults, assures, quotation, first

Table II: Six Polysemous Words			
English	French	sense	N
duty	droit	tax	1114
	devoir	obligation	691
drug	médicament	medical	2992
	drogue	illicit	855
land	terre	property	1022
	pays	country	386
language	langue	medium	3710
	langage	style	170
position	position	place	5177
	poste	job	577
sentence	peine	judicial	296
	phrase	grammatical	148

Table V: Sample Concordances of *duty* (split into two senses)

Sense	Examples (from Canadian Hansards)
tax	fewer cases of companies paying >duty< and then claiming a refund
	and impose a countervailing >duty< of 29,1 per cent on candian exports of
	the united states imposed a >duty< on canadian saltfish last year
obligation	it is my honour and >duty< to present a petition duly approved
	working well beyond the call of >duty< ? SENT i know what time they start
	in addition , it is my >duty< to present the government 's comments

$$\frac{L(sense_1)}{L(sense_2)} \approx \prod_{tok\ in\ context} \frac{Pr(tok \mid sense_1)}{Pr(tok \mid sense_2)}$$

Metaphor: Classic Hard Problem in NLP

- Stereotypes: Get Smart
- Considerable literature
  - Carbonell (1980) https://aclanthology.org/P80-1004
  - Fass & Wilks (1983) https://aclanthology.org/J83-3004
  - Martin (1990) *A Computational Model of Metaphor Interpretation*
  - Hobbs (1992) *Metaphor and Abduction*
  - Gedigian et al (2006) https://aclanthology.org/W06-3506
  - Krishnakumaran and Zhu (2007) https://aclanthology.org/W07-0103
  - Lakoff (2008) *Women, Fire and Dangerous Things*
  - Lakoff and Johnson (2008) *Metaphors to Live By*
  - Shutova (2010) https://aclanthology.org/P10-1071
  - Mohammad et al (2016) https://aclanthology.org/S16-2003

- *cover all the bases*
- *drop the ball*
- *dunk*
- *fumble*
- *get on base*
- *hit a home run*
- *out in left field*
- *punt*
- *ragging the puck*
- *run out the clock*
- *sticky wicket*
- *strike out*

Repositories

- [HuggingFace]
- [LDC] (Linguistic Data Consortium)
- [NLTK]

WordNet: An Example of an Ontology

https://en.wikipedia.org/wiki/WordNet

Knowledge structure [edit]

Both nouns and verbs are organized into hierarchies, defined by hypernym or *IS A* relationships. For instance, one sense of the word *dog* is found following hypernym hierarchy; the words at the same level represent synset members. Each set of synonyms has a unique index.

- dog, domestic dog, Canis familiaris
  - canine, canid
    - carnivore
      - placental, placental mammal, eutherian, eutherian mammal
        - mammal
          - vertebrate, craniate
            - chordate
              - animal, animate being, beast, brute, creature, fauna
                - ...

https://aclanthology.org/2021.emnlp-main.501.pdf

Relation	Edges	Inverse	Edges
hypernyms	37,221	hyponyms	37,221
derivationally related forms	31,867		
member meronym	7928	member holonum	7928
has part	5142	part of	5148
synset domain topic of	3335	member of domain topic	3341
instance hypernym	3150	instance hyponym	3150
also see	1396		
verb group	1220		
member of domain region	983	synset domain region of	982
member of domain usage	675	synset domain usage of	669
similar to	86		

Table 2: 18 Relations in WN18. By construction, many of these relations have inverses (with similar counts).

MeSH

HIERARCHY TREE VIEW

MeSH | ICD-10

Anatomy [A]	A
Organisms [B]	B
Diseases [C]	C
Neoplasms *1 indication for 3418 drugs (688 approved, 2730 experimental)*	C04
Neoplasms by Site *1 indication for 48 drugs (30 approved, 18 experimental)*	C04.588
Abdominal Neoplasms *1 indication for 24 drugs (22 approved, 2 experimental)*	C04.588.033
Anal Gland Neoplasms	C04.588.083
Bone Neoplasms *1 indication for 41 drugs (29 approved, 12 experimental)*	C04.588.149
Breast Neoplasms *1 indication for 1583 drugs (514 approved, 1069 experimental)*	C04.588.180
Breast Carcinoma In Situ *1 indication for 12 drugs (11 approved, 1 experimental)*	C04.588.180.130
Breast Neoplasms, Male *1 indication for 100 drugs (59 approved, 41 experimental)*	C04.588.180.260
Carcinoma, Ductal, Breast *1 indication for 12 drugs (8 approved, 4 experimental)*	C04.588.180.390
Carcinoma, Lobular *1 indication for 3 approved drugs*	C04.588.180.437
Hereditary Breast and Ovarian Cancer Syndrome *1 indication for 5 drugs (3 approved, 2 experimental)*	C04.588.180.483
Inflammatory Breast Neoplasms *1 indication for 44 drugs (36 approved, 8 experimental)*	C04.588.180.576
Triple Negative Breast Neoplasms *1 indication for 294 drugs (89 approved, 205 experimental)*	C04.588.180.788
Unilateral Breast Neoplasms	C04.588.180.800
Digestive System Neoplasms *1 indication for 60 drugs (33 approved, 27 experimental)*	C04.588.274
Endocrine Gland Neoplasms *1 indication for 16 drugs (11 approved, 5 experimental)*	C04.588.322
Eye Neoplasms *1 indication for 6 drugs (4 approved, 2 experimental)*	C04.588.364
Head and Neck Neoplasms *1 indication for 496 drugs (208 approved, 288 experimental)*	C04.588.443
Hematologic Neoplasms *1 indication for 252 drugs (125 approved, 127 experimental)*	C04.588.448

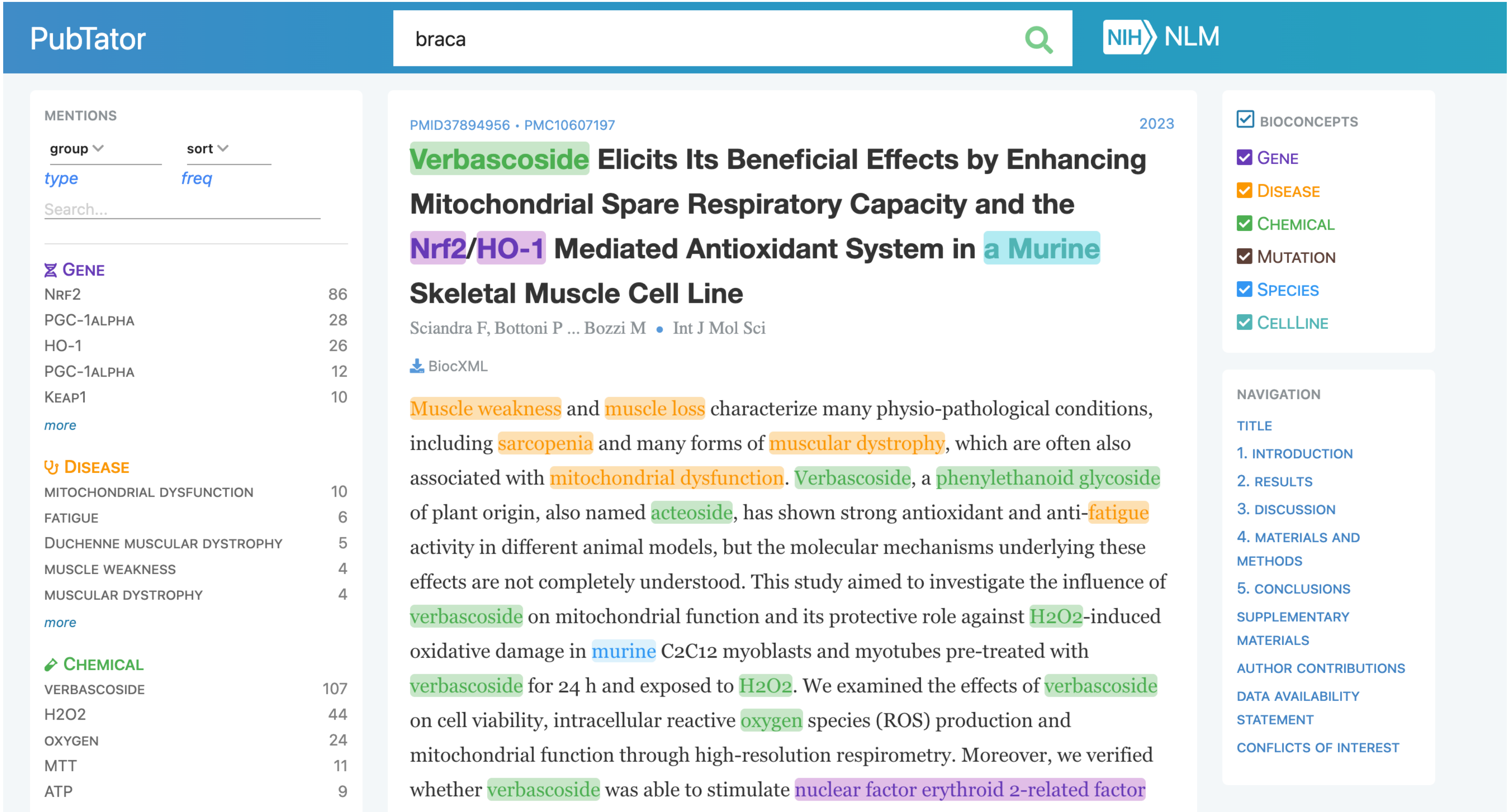
JM23


```
bass³, basso (an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => organism, being
                  => living thing, animate thing,
                     => whole, unit
                        => object, physical object
                           => physical entity
                              => entity
bass⁷ (member with the lowest range of a family of instruments)
=> musical instrument, instrument
   => device
      => instrumentality, instrumentation
         => artifact, artefact
            => whole, unit
               => object, physical object
                  => physical entity
                     => entity
```

Figure 23.5 Hyponymy chains for two separate senses of the lemma *bass*. Note that the chains are completely distinct, only converging at the very abstract level *whole, unit*.

Tasks

- Word Sense Disambiguation
  - *bank* → ``money’’ bank vs. ``river’’ bank
- NER (Named Entity Recognition)
  - Find spans
- Linking: Add hypertext links from texts to resources
  - Wikipedia
  - Pubtator
- Co-reference
  - Which nouns refer to which nouns?
  - Pronoun resolution → Winograd Schema
- Stance, Sentiment, Synonyms vs. Antonyms, Negation


PubTator
braca
NIH NLM
MENTIONS
group
sort
type
freq
Search...
GENE
NRF2 86
PGC-1ALPHA 28
HO-1 26
PGC-1ALPHA 12
KEAP1 10
more
DISEASE
MITOCHONDRIAL DYSFUNCTION 10
FATIGUE 6
DUCHENNE MUSCULAR DYSTROPHY 5
MUSCLE WEAKNESS 4
MUSCULAR DYSTROPHY 4
more
CHEMICAL
VERBASCOSIDE 107
H2O2 44
OXYGEN 24
MTT 11
ATP 9
PMID37894956 • PMC10607197
2023
Verbascoside Elicits Its Beneficial Effects by Enhancing Mitochondrial Spare Respiratory Capacity and the Nrf2/HO-1 Mediated Antioxidant System in a Murine Skeletal Muscle Cell Line
Sciandra F, Bottoni P ... Bozzi M • Int J Mol Sci
BiocXML
Muscle weakness and muscle loss characterize many physio-pathological conditions, including sarcopenia and many forms of muscular dystrophy, which are often also associated with mitochondrial dysfunction. Verbascoside, a phenylethanoid glycoside of plant origin, also named acteoside, has shown strong antioxidant and anti-fatigue activity in different animal models, but the molecular mechanisms underlying these effects are not completely understood. This study aimed to investigate the influence of verbascoside on mitochondrial function and its protective role against H2O2-induced oxidative damage in murine C2C12 myoblasts and myotubes pre-treated with verbascoside for 24 h and exposed to H2O2. We examined the effects of verbascoside on cell viability, intracellular reactive oxygen species (ROS) production and mitochondrial function through high-resolution respirometry. Moreover, we verified whether verbascoside was able to stimulate nuclear factor erythroid 2-related factor
BIOCONCEPTS
GENE
DISEASE
CHEMICAL
MUTATION
SPECIES
CELLLINE
NAVIGATION
TITLE
1. INTRODUCTION
2. RESULTS
3. DISCUSSION
4. MATERIALS AND METHODS
5. CONCLUSIONS
SUPPLEMENTARY MATERIALS
AUTHOR CONTRIBUTIONS
DATA AVAILABILITY STATEMENT
CONFLICTS OF INTEREST

Pubtator

Winograd Schema (GLUE WNLI)

- The trophy doesn't fit in the brown suitcase
  - because it is too large/small.
- What is too large?
  - A. The trophy
  - B. The suitcase

Task	Metric	Result	Training time
CoLA	Matthews corr	56.53	3:17
SST-2	Accuracy	92.32	26:06
MRPC	F1/Accuracy	88.85/84.07	2:21
STS-B	Pearson/Spearman corr.	88.64/88.48	2:13
QQP	Accuracy/F1	90.71/87.49	2:22:26
MNLI	Matched acc./Mismatched acc.	83.91/84.10	2:35:23
QNLI	Accuracy	90.66	40:57
RTE	Accuracy	65.70	57
WNLI	Accuracy	56.34	24

Not much better than chance

Table 1. Time line of the Winograd Schema Challenge.

1972:	Winograd's (1972) thesis introduces the original example.
2010:	Levesque [47] proposes the Winograd Schema Challenge.
2010–2011:	The initial corpus of Winograd schemas is created [50].
2014:	Levesque's Research Excellence talk "On our best behavior" [48].
2016:	The Winograd Schema Challenge is run at IJCAI-16. No systems do much better than chance [16].
2018:	WNLI is incorporated in the GLUE set of benchmarks. BERT-based systems do no better than most-frequent-class guessing [91].
2019, May:	Kocijan et al. [43] achieve 72.5% accuracy on WSC273 using pretraining.
2019, June:	Liu et al. [56] achieve 89.0% on WNLI.
2019, November:	Sakaguchi et al. [77] achieve 90.1% on WSC273.

from: https://doi.org/10.1016/j.artint.2023.103971

Winograd Schema (GLUE WNLI)

A Surprisingly Robust Trick for the Winograd Schema Challenge

Vid Kocijan[1], Ana-Maria Creţu[2], Oana-Maria Camburu[1,3], Yordan Yordanov[1], Thomas Lukasiewicz[1,3]
[1]University of Oxford
[2]Imperial College London
[3]Alan Turing Institute, London
firstname.lastname@cs.ox.ac.uk, a.cretu@imperial.ac.uk

Abstract

The Winograd Schema Challenge (WSC) dataset WSC273 and its inference counterpart WNLI are popular benchmarks for natural language understanding and commonsense reasoning. In this paper, we show that the performance of three language models on WSC273 consistently and robustly improves when fine-tuned on a similar pronoun disambiguation problem dataset (denoted WSCR). We additionally generate a large unsupervised WSC-like dataset. By fine-tuning the BERT language model both on the introduced and on the WSCR dataset, we achieve overall accuracies of 72.5% and 74.7% on WSC273 and WNLI, improving the previous state-of-the-art solutions by 8.8% and 9.6%, respectively. Furthermore, our fine-tuned models are also consistently more accurate on the "complex" subsets of WSC273, introduced by Trichelair et al. (2018).

to the small existing datasets making it difficult to train neural networks directly on the task.

Neural networks have proven highly effective in natural language processing (NLP) tasks, outperforming other machine learning methods and even matching human performance (Hassan et al., 2018; Nangia and Bowman, 2018). However, supervised models require many per-task annotated training examples for a good performance. For tasks with scarce data, transfer learning is often applied (Howard and Ruder, 2018; Johnson and Zhang, 2017), i.e., a model that is already trained on one NLP task is used as a starting point for other NLP tasks.

A common approach to transfer learning in NLP is to train a language model (LM) on large amounts of unsupervised text (Howard and Ruder, 2018) and use it, with or without further fine-tuning, to solve other downstream tasks. Building on top of a LM has proven to be very suc-

View PDF

Artificial Intelligence
Available online 11 July 2023, 103971
In Press, Corrected Proof   What's this?

The defeat of the Winograd Schema Challenge

Vid Kocijan [a 1], Ernest Davis [b], Thomas Lukasiewicz [c d], Gary Marcus [e], Leora Morgenstern [f]

Show more

Add to Mendeley   Share   Cite

https://doi.org/10.1016/j.artint.2023.103971   Get rights and content

Abstract

The Winograd Schema Challenge—a set of twin sentences involving pronoun reference disambiguation that seem to require the use of commonsense knowledge—was proposed by Hector Levesque in 2011. By 2019, a number of AI systems, based on large pre-trained transformer-based language models and fine-tuned on these kinds of problems, achieved better than 90% accuracy. In this paper, we review the history of the Winograd Schema Challenge and discuss the lasting contributions of the flurry of research that has taken place on the WSC in the last decade. We discuss the significance of various datasets developed for WSC, and the research community's deeper understanding of the role of surrogate tasks in assessing the intelligence of an AI system.

Keywords

Commonsense reasoning; Winograd Schema Challenge

Training on Lexical Resources

https://aclanthology.org/2022.lrec-1.676.pdf

$$rel \sim w_1 + w_2 \quad (1)$$

The fine-tuning code is very simple. We modified an example from HuggingFace[2] in straightforward ways.[3] This code takes a pretrained net as input, and a set of triples, and outputs a fine-tuned net.

$text_1$	$text_2$	y_1	y_2
good	bad	-3.95	4.54
bad	evil	4.44	-5.00
good	benevolent	4.43	-5.05
bad	benevolent	-3.44	4.16
good	terrorist	-3.43	4.10
bad	terrorist	4.48	-5.10

Table 1: Inference: synonymy iff $y_1 > y_2$

$$y \sim text_1 + text_2 \quad (2)$$

This notation is inspired by general linear models in R[6] (Guisan et al., 2002). We will start with binary classification (logistic regression). Later, classification will be replaced with regression when we consider VAD (Valance, Arousal and Dominance) distances in §5.

$text_1$	$text_2$	y_1	y_2
freedom fighter	good	2.33	-2.56
freedom fighter	bad	-1.50	2.19
white supremacist	good	-2.05	2.91
white supremacist	bad	1.67	-1.61

Table 2: Mutiword Expressions (MWEs)

[scipy dendrogram]

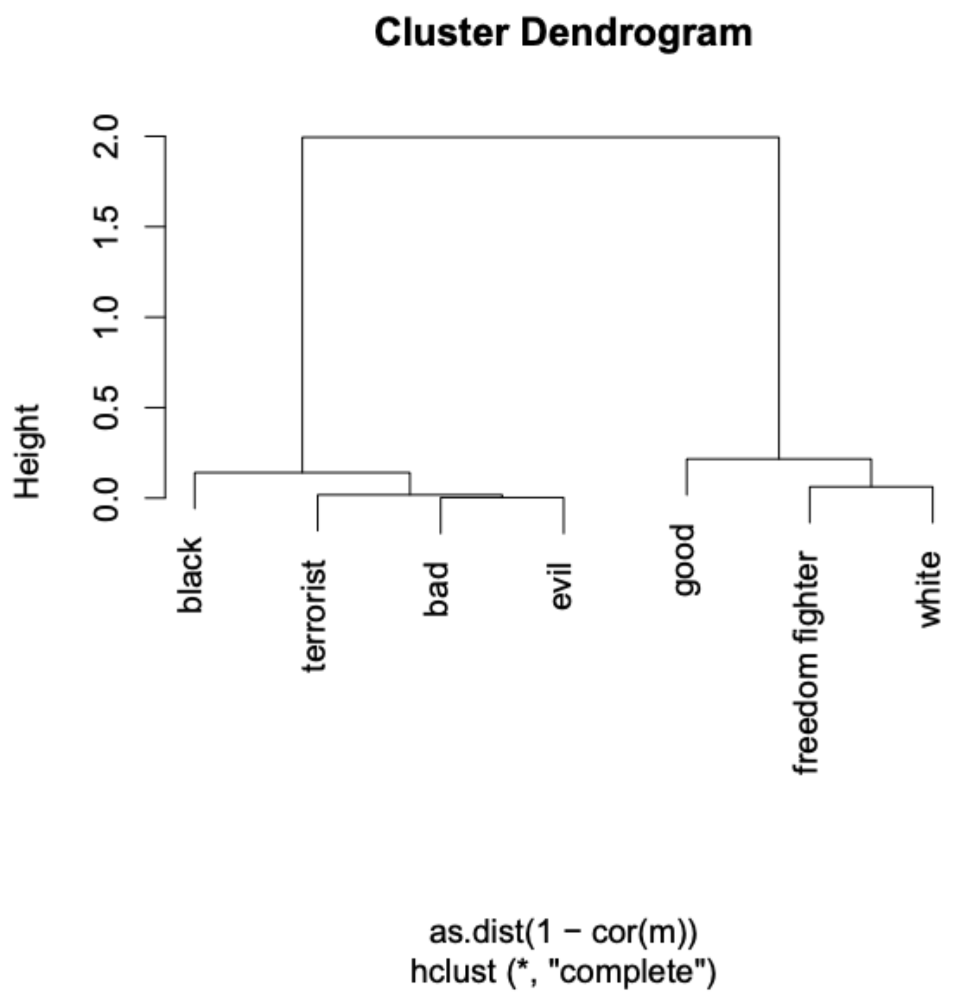


Figure 1: Clustering of correlations in Table 8 (bottom), illustrating biases in model.

https://en.wikipedia.org/wiki/Hierarchical_clustering

https://en.wikipedia.org/wiki/K-means_clustering

Clustering in Scikit-Learn

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters or distance threshold	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters or distance threshold, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes	Distances between nearest points
OPTICS	minimum cluster membership	Very large `n_samples`, large `n_clusters`	Non-flat geometry, uneven cluster sizes, variable cluster density	Distances between points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	branching factor, threshold, optional global clusterer.	Large `n_clusters` and `n_samples`	Large dataset, outlier removal, data reduction.	Euclidean distance between points

https://scikit-learn.org/stable/modules/clustering.html

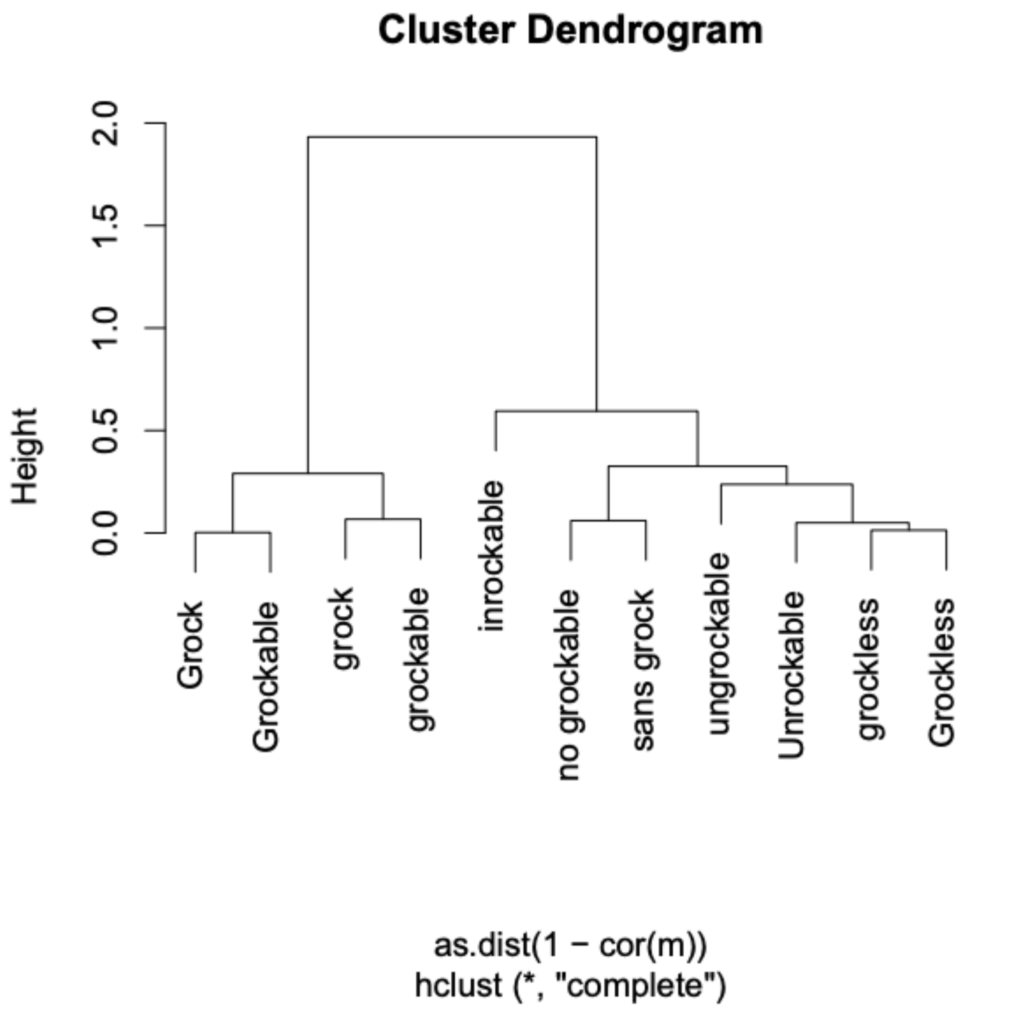


Figure 3: Clustering of morphological variants and translations of an out-of-vocabulary (OOV) word: *grock*. Base model: bert-base-multilingual-cased.

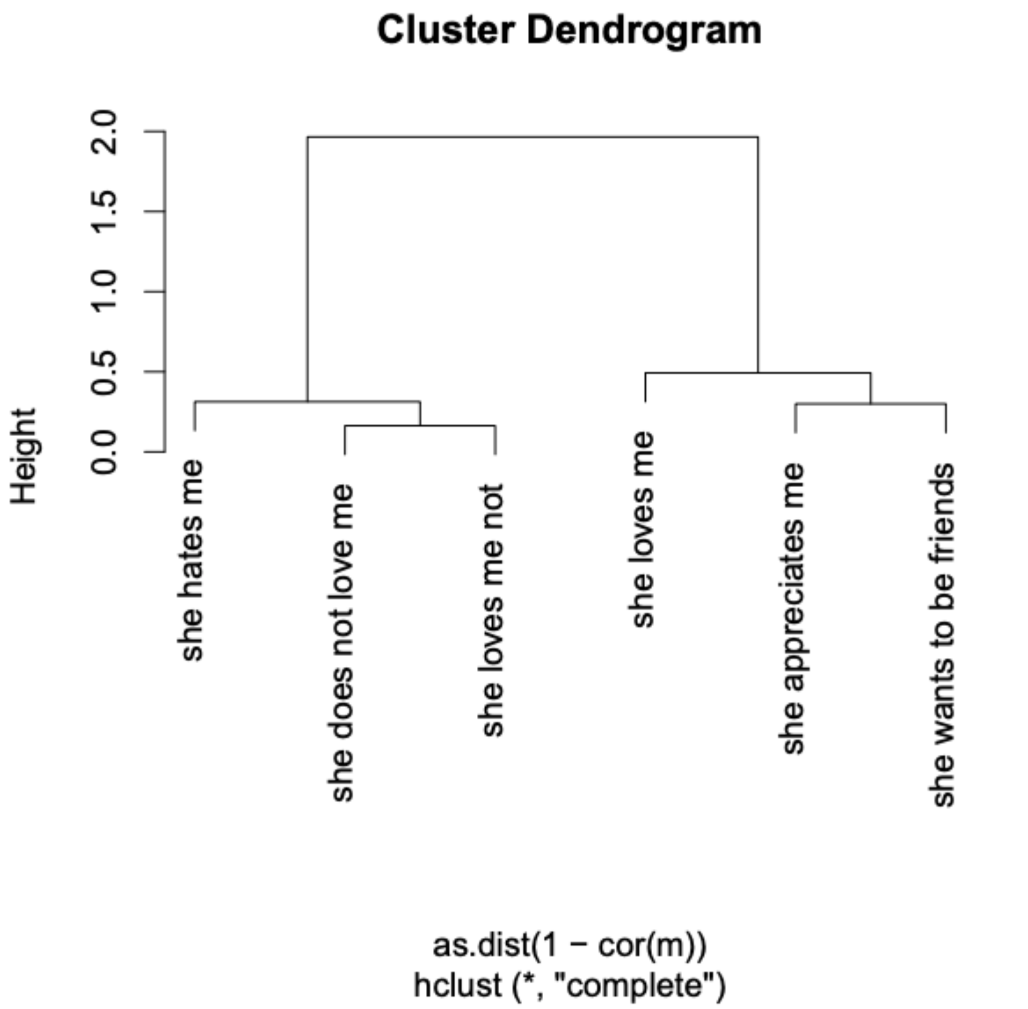


Figure 4: Clustering of correlations of logits of all pairs of six sentences.

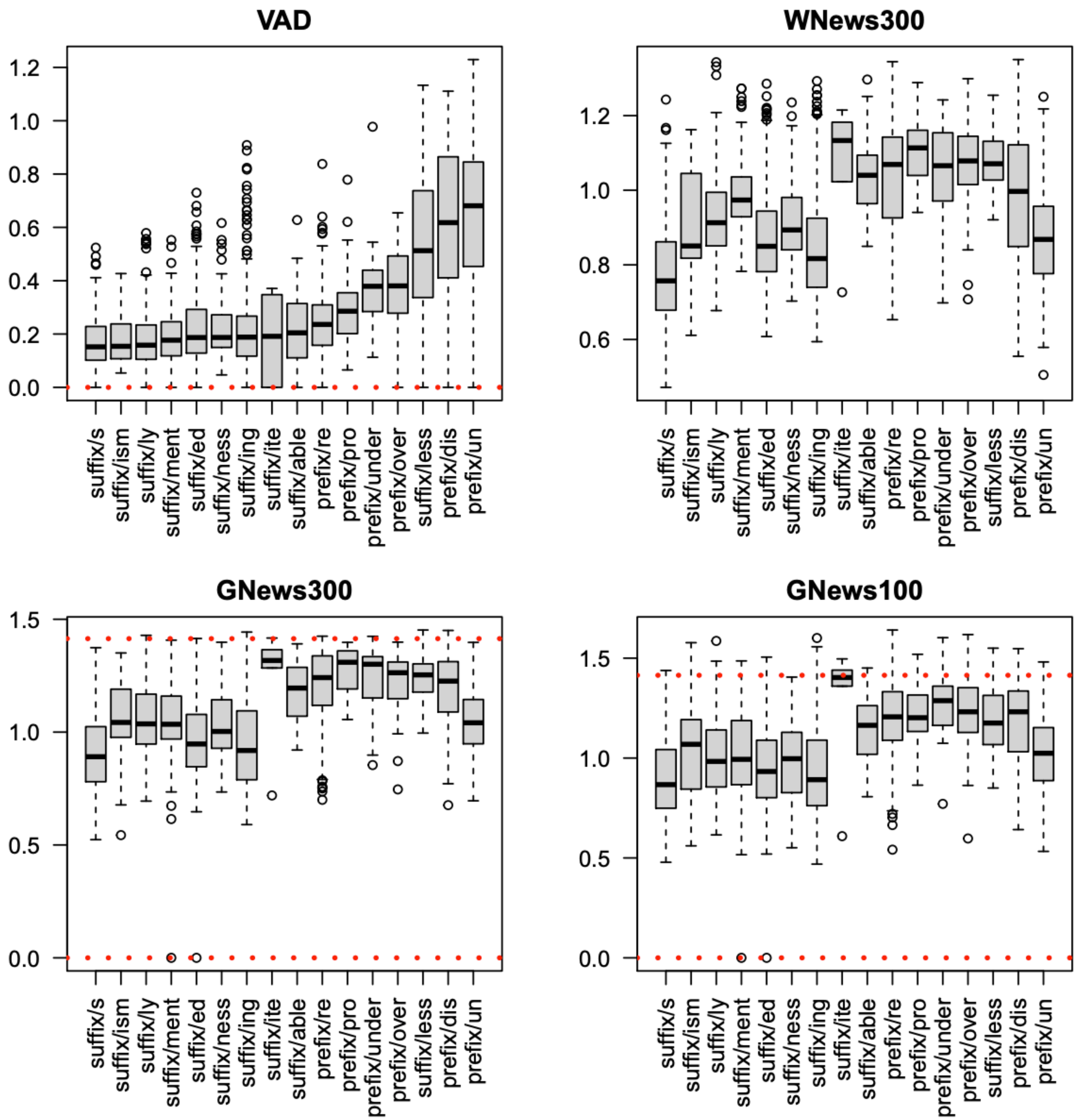
https://saifmohammad.com/WebPages/nrc-vad.html
https://saifmohammad.com/WebDocs/VAD-talk.pdf

word	Val	Arousal	Dom	Dist
open	0.620	0.480	0.569	0.00
unfold	0.612	0.510	0.520	0.06
reopen	0.656	0.528	0.568	0.06
close	0.292	0.260	0.263	0.50
closed	0.240	0.164	0.318	0.55
undecided	0.286	0.433	0.127	0.56

Table 12: Words above the double line are near *open*. The last column is the Euclidean distance to *open*.

Church et al.,
Emerging trends:
General fine-tuning (gft)
Natural Language Engineering, 28(4), 519-535.
doi:10.1017/S1351324922000237

–Data	–eqn
H:glue,cola	classify: label ~ sentence
H:glue,sst2	classify: label ~ sentence
H:glue,wnli	classify: label ~ sentence
H:glue,mrpc	classify: label ~ sentence1 + sentence2
H:glue,rte	classify: label ~ sentence1 + sentence2
H:glue,qnli	classify: label ~ question + sentence
H:glue,qqp	classify: label ~ question1 + question2
H:glue,sstb	regress: label ~ sentence1 + sentence2
H:glue,mnli	classify: label ~ premise + hypothesis
H:squad	classify_spans: answers ~ question + context
H:squad_v2	classify_spans: answers ~ question + context
H:tweet_eval,hate	classify: label ~ text
H:conll2003	classify_tokens: pos_tags ~ tokens
H:conll2003	classify_tokens: ner_tags ~ tokens
H:conll2003	classify_tokens: chunk_tags ~ tokens
H:timit_asr	ctc: text ~ audio
H:librispeech_asr	ctc: text ~ audio
C:$gft/datasets/VAD/VAD	regress: Valence + Arousal + Dominance ~ Word



VAD
WNews300
GNews300
GNews100
suffix/s
suffix/ism
suffix/ly
suffix/ment
suffix/ed
suffix/ness
suffix/ing
suffix/ite
suffix/able
prefix/re
prefix/pro
prefix/under
prefix/over
suffix/less
prefix/dis
prefix/un

Lexical Resources

International Conference on Language Resources and Evaluation (LREC)

- Corpora
  - Non-parallel:
    - Brown, Penn Treebank, Wikitext
  - Parallel:
    - Hansards, Europarl
- Ontologies
  - WordNet
  - MeSH (Medical Subject Headings)
- Dictionaries
  - CMU Dict
- Thesaurus
  - Roget's
  - Synonyms and Antonyms
  - NRC-VAD
- Knowledge Graphs
  - <head, relation, tail>
  - FreeBase (FB15k)
  - WordNet (WN18RR)

Example of Parallel Corpus
https://youtu.be/1jeDPcWEYX0?t=80

	A	B	C
1	English	Spanish	French
2	"What's it going to be then, eh?"	–¿Y ahora qué pasa, eh?	— Bon, alors ça sera quoi, hein ?
3	There was me, that is Alex, and my three droogs, that is Pete, Georgie, and Dim, Dim being really dim, and we sat in the Korova Milkbar making up our rassoodocks what to do with the evening, a flip dark chill winter bastard though dry.	Estábamos yo, Alex, y mis tres drugos, Pete, Georgie y el Lerdo, que realmente era lerdo, sentados en el bar lácteo Korova, exprimiéndonos los rasudoques y decidiendo qué podríamos hacer esa noche, en un invierno oscuro, helado y bastardo aunque seco.	Il y avait moi, autrement dit Alex, et mes trois drougs, autrement dit Pierrot, Jo et Momo, vraiment momo le Momo, et on était assis au Korova Milkbar à se creuser le rassoudok pour savoir ce qu'on ferait de la soirée, – une putain de soirée d'hiver, branque, noire et glaciale, mais sans eau.
4	The Korova Milkbar was a milk-plus mesto, and you may, O my brothers, have forgotten what these mestos were like, things changing so skorry these days and everybody very quick to forget, newspapers not being read much neither.	El bar lácteo Korova era un mesto donde servían leche-plus, y quizás ustedes, oh hermanos míos, han olvidado cómo eran esos mestos, pues las cosas cambian tan scorro en estos días, y todos olvidan tan rápido, aparte de que tampoco se leen mucho los diarios.	Le Korova Milkbar, c'était un de ces messtots où on servait du lait gonflé, et peut-être avez-vous oublié, Ô mes frères, à quoi ressemblait ce genre de messtot, tellement les choses changent zoum par les temps qui courent et tellement on a vite fait d'oublier, vu aussi qu'on ne lit plus guère les journaux.
5	Well, what they sold there was milk plus something else.	Bueno, allí vendían leche con algo más.	Bref ce qu'on y vendait c'était du lait gonflé à autre chose.
6	They had no license for selling liquor, but there was no law yet against prodding some of the new veshches which they used to put into the old moloko, so you could peet it with vellocet or synthemesc or drencrom or one or two other veshches which would give you a nice quiet horrorshow fifteen minutes admiring Bog And All His Holy Angels and Saints in your left shoe with lights bursting all over your mozg.	No tenían permiso para vender alcohol, pero en ese tiempo no había ninguna ley que prohibiese las nuevas vesches que acostumbraban meter en el viejo moloco, de modo que se podía pitearlo con velocet o synthemesco o drencrom o una o dos vesches más que te daban unos buenos, tranquilos y joroschós quince minutos admirando a Bogo y el Coro Celestial de Ángeles y Santos en el zapato izquierdo, mientras las luces te estallaban en el mosco.	Le Korova n'avait pas de licence pour la vente de l'alcool, mais il n'existait pas encore de loi interdisant d'injecter de ces nouvelles vesches qu'on mettait à l'époque dans le moloko des familles, ce qui faisait qu'on pouvait le drinker avec de la vélocette, du synthémesc ou du methcath, ou une ou deux autres vesches, et s'offrir quinze gentilles minutes pépère tzarrible à mirer Gogre et Tous Ses Anges et Ses Saints dans son soulier gauche, le mozg plein à péter de lumières.

Applications for Parallel Corpora

- Machine Translation
- Word Sense Disambiguation

$$\prod_{w \text{ in doc}} \frac{Pr(w|rel)}{Pr(w|irrel)}$$ Information Retreival (IR)

$$\prod_{w \text{ in doc}} \frac{Pr(w|author_1)}{Pr(w|author_2)}$$ Author Identification

In the sense disambiguation application, the 100-word context surrounding instances of a polysemous word (e.g., *sentence*) are treated very much like a document.[1]

$$\prod_{w \text{ in context}} \frac{Pr(w|sense_1)}{Pr(w|sense_2)}$$ Sense Disambiguation

$$\prod_{w \text{ in context}} Pr(w|Roget\ Category_i)$$

The program can also be run in a mode where it takes unrestricted text as input and tags each word with its most likely Roget Category. Some results for the word *crane* are presented below, showing that the program can be used to sort a concordance by sense.

Input		Output
Treadmills attached to	*cranes* were used to lift heavy	TOOLS
for supplying power for	*cranes* , hoists , and lifts	TOOLS
Above this height , a tower	*crane* is often used .SB This	TOOLS
elaborate courtship rituals	*cranes* build a nest of vegetation	ANIMAL
are more closely related to	*cranes* and rails .SB They range	ANIMAL
low trees .PP At least five	*crane* species are in danger of	ANIMAL

https://aclanthology.org/P92-1032.pdf

Table 4
A bilingual concordance.

bank/banque ("money" sense)	
¿ it could also be a place where we would have a	bank of experts. SENT i know several people who a
ftre le lieu oü se retrouverait une espèce de	banque d' experts. SENT je connais plusieurs pers
f finance (mr. wilson) and the governor of the	bank of canada have frequently on behalf of the ca
es finances (m . wilson) et le gouverneur de la	banque du canada ont fréquemment utilisé au co
reduced by over 800 per cent in one week through	bank action. SENT there was a haberdasher who wou
us de 800 p. 100 en une semaine à cause d'une	banque. SENT voilà un chemisier qui aurait appr
bank/banc ("place" sense)	
h a forum. SENT such was the case in the georges	bank issue which was settled between canada and th
entre les états-unis et le canada à propos du	banc de george. SENT c'est dans le but de ré
han i did. SENT he said the nose and tail of the	bank were surrendered by this government. SENT th
gouvernement avait cédé les extrémités du	banc. SENT en fait, lors des négociations de l
he fishing privileges on the nose and tail of the	bank went down the tube before we even negotiated
les privilèges de pêche aux extrémités du	banc ont été liquidés avant même qu' on ai

https://aclanthology.org/J93-1004.pdf

Co-Reference

- Slides from last term
- JM26
- Two Noriegas