University of Stuttgart
Germany

Automatic Dictionary Induction

October 22, 2024

Dominik Schlechtweg, Wei Zhao
University of Stuttgart, University of Aberdeen

Problem

- updating dictionaries is an **important**, **never-ending** and **resource-intensive** task
- includes:
  - finding neologisms in a reference corpus,
  - finding non-recorded senses in a reference corpus,
  - generating definitions from usages with non-recorded senses or neologisms,
  - finding illustrative example usages for senses,
  - ...

NSD through Word Sense Disambiguation

usage: . . . and though he saw her within reach of his **arm**, yet the light of her eyes seemed as far off. . .

senses: “a human limb”; “a weapon”; “any projection that is thought to resemble a human arm”;. . .

Task: Is the usage described by any sense?

Data: WordNet, Swedish Academy Dictionary, Leipzig corpora, Historical corpora

Solution: Lautenschlager et al. (2024): Use semantic Word-in-Context encoder with cosine similarity to compare usage to senses.

Results

	English	Swedish
Usages	322	927
Non-recorded	14%	65%
Random Baseline	10%	17%

Examples

usage: There should be some things that can be done in the short term, but in terms of developing the **pipeline** further on coaching and executive positions, that would take a longer period of time.

senses: "a pipe used to transport liquids or gases"; "gossip spread by spoken communication"

OED senses: "figurative. A channel of supply, information, communication, etc.; a means of ready access"; ...

Examples

usage: No wonder he's up there getting big **baked**.

senses: "cook and make edible by putting in a hot oven"; "prepare with dry heat in an oven"; "heat by a natural force"; "be very hot, due to hot weather or exposure to the sun"; "dried out by heat or excessive exposure to sunlight"; "(bread and pastries) cooked by dry heat (as in an oven)"

OED senses: "Intoxicated by alcohol; drunk. Also (now chiefly): intoxicated by a recreational drug, esp. marijuana; high"; ...

Examples

usage: You seem to intend a eulogy, yet leave out whatever was noblest in her, and **blacken** while you mean to praise.

senses: "make or become black"; "burn slightly and superficially so as to affect color"

OED senses: "figurative. To defame (a person), to damage (a person's reputation, name, etc.).";...

Analysis

- major problem: multi-word expressions

  usage: I'm at that age where many of my friends are having children, and a central topic of conversation whenever we're together **revolves** around creating the almost scientifically set schedule for their babies.

  senses: "turn on or around an axis or a center"; "move in an orbit"; "cause to move by turning over or in a circular manner of as if on an axis"

NSD through Word Sense Induction

usage1: . . . and though he saw her within reach of his **arm**,. . .

usage2: . . . and taking a knife from her pocket, she opened a vein in her little **arm**.

usage3: . . . had been heavily taxed to pay for the **arms**, ammunition;

senses: "a human limb"; "a weapon"; "any projection that is thought to resemble a human arm";. . .

Task: Does the number of senses in the corpus correspond to the number of senses in the dictionary?

Data: OED, WordNet, DWDS, Swedish Academy Dictionary, Leipzig corpora

Solution: Sander et al. (2024); Sköldberg et al. (2024): Word-in-Context encoder + Clustering algorithm, DURel tool

Results

	Sampled	>1 cluster	New sense(s)
OED	103	43%	47%
WordNet	100	40%	10%
DWDS	108	43%	37%

- Monosemous headwords
- Headwords are from OED updates 2015-2024; uses are in Leipzig Corpora (10M).
  - 44 headwords have 1+ clusters.
  - 21 headwords that we discover novel senses, but some appear only a few times.
  - 23 faulty clusters
  - Why 47%?
  - Polysemous words?

Examples

Headword	OED sense	New sense(s)	Percent	in-ref-dicts
20 new senses are proper nouns:				
AQI	Air Quality Index	al-Qaeda in Iraq	4%	N
Assisi	A style of embroidery	A city in Italy	16%	Y
Ealing	A film made by ES	A district in London	8%	Y
Marco Polo	A tag game	A person/motive name	96%	Y
1 new sense is medical use:				
Broken heart	A state of grief	Medical use for heart attack	8%	Y

- Reference dictionaries: Collins, Dictionary.com, Cambridge, Merriam-Webster
- When to draft a new sense?
  - Frequency
  - Recorded in other dictionaries
  - OED standards

SDG through finetuning LLMs

usage1: . . . and though he saw her within reach of his **arm**,. . .

usage2: . . . and taking a knife from her pocket, she opened a vein in her little **arm**.

Task: What is an adequate dictionary definition for the sense expressed by the usages?

Data: OED, WordNet, Wiktionary, Bokmålsordboka

Solution: Kutuzov et al. (2024): Fine-tune an LLM (mT0-xl) on usage-definition pairs.

Results

Model	Accuracy
DefGen	.69
Lesk	.22

Examples

prompt: . . . The representative of the United States Government speaks about seeds for rice with, increased yield which the United States will send to Vietnam," Mr. Thuy said, "but even at this moment American troops and American **planes** are attacking our harvest, are now destroying the economic infrastructure of South Vietnam." What is the definition of **plane**?

Definition by mT0: An aircraft.

Definition by GPT-3.5-turbo: An aircraft that is used for military purposes, typically for combat or warfare.

Definition by GPT-o1-preview: An airplane; a powered flying vehicle with fixed wings and a weight greater than that of the air it displaces, used for transportation or military purposes.

Definition in OED: An aeroplane.

Examples

prompt: By the adoption of stationary powers, on inclined **planes**, in surmounting some of the steeper declivities, it is supposed the distance might be reduced to something less than a hundred and ninety miles, without any considerable increase of the elevation to be passed. What is the definition of **plane**?

Definition by mT0: A flat surface, without slope, tilts, or indentations; a level surface.

Definition by GPT-3.5-turbo: A *flat or level surface*, often used to describe a sloping surface like *an inclined plane*.

Definition by GPT-o1-preview: An *inclined plane*, which is a flat, sloping surface or engineered gradient used to move loads or vehicles up or down steep slopes. It is a simple *machine* that reduces the effort needed to raise or lower objects by allowing them to be moved over a longer distance at a gentler angle.

Definition in OED: A flat or level surface of a material body. Also: the flatness of a material surface.

Analysis

- major problem: too general definitions

  prompt: It follows, that if the point in space be upon the horizontal **plane**, its distance with regard to this last named **plane** will be zero or nothing, and the vertical A a will be zero also. What is the definition of **plane**?

  Definition by mT0: A surface without outline; a flat or almost flat surface.

  Definition in OED: A flat geometrical surface which has the property that every straight line joining any two points of the surface lies wholly in the surface.

Automatic Prioritization Of DICTionary Update candidateS (APODICTUS)

- **Background**: The Oxford English Dictionary (OED, 2009) stores candidate entries for new words and senses in an internal database, which has currently roughly 60,000 entries. The prioritization of these entries is currently done with only limited automation.
- **Aim**: Automatically assign a prioritization score to each entry in the database.
- **Benefit**: More systematic workflow in dictionary updates.
- **Output**: Mapping of each database entry to a prioritization score with further information such as occurrence and predicted sense frequency.
- **Modeling**: The score for each entry will be assigned based on an automatic comparison between the headword entry in the public OED, a large reference corpus, and information from the database entry such as the tentative sense description (e.g. Kokosinskii, Kuklin, & Arefyev, 2024; Lautenschlager et al., 2024; Scarlini, Pasini, & Navigli, 2020). Database entries will be prioritized according to the amount of corpus evidence found for the suggested new sense of the headword.
- **Supervisors**: Dominik Schlechtweg (Stuttgart), Wei Zhao (Aberdeen).

References I

Kokosinskii, D., Kuklin, M., & Arefyev, N. (2024, aug). Deep-change at AXOLOTL-24: Orchestrating WSD and WSI models for semantic change modeling. In N. Tahmasebi et al. (Eds.), *Proceedings of the 5th workshop on computational approaches to historical language change* (pp. 168–179). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.lchange-1.16 doi: 10.18653/v1/2024.lchange-1.16

Kutuzov, A., Fedorova, M., Schlechtweg, D., & Arefyev, N. (2024, May). Enriching word usage graphs with cluster definitions. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 6189–6198). Torino, Italia: ELRA and ICCL. Retrieved from https://aclanthology.org/2024.lrec-main.546

Lautenschlager, J., Hengchen, S., & Schlechtweg, D. (2024). *Detection of non-recorded word senses in English and Swedish*. Retrieved from https://arxiv.org/abs/2403.02285

OED. (2009). *Oxford english dictionary*. Online: Oxford University Press.

Sander, P., Hengchen, S., Zhao, W., Ma, X., Sköldberg, E., Virk, S. M., & Schlechtweg, D. (2024). The DURel Annotation Tool: Using fine-tuned LLMs to discover non-recorded senses in multiple languages. In *Workshop on Large Language Models and Lexicography at 21st EURALEX International Congress Lexicography and Semantics*.

Scarlini, B., Pasini, T., & Navigli, R. (2020, Apr.). SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8758–8765. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/6402 doi: 10.1609/aaai.v34i05.6402

Sköldberg, E., Virk, S. M., Sander, P., Hengchen, S., & Schlechtweg, D. (2024). Revealing semantic variation in Swedish using computational models of semantic proximity: results from lexicographical experiments. In *21st EURALEX International Congress Lexicography and Semantics*.