Basic Level Detection:

Learning from corpus characteristics and synthetic features

Haochen Wang (CS, VU & UvA Joint Degree)

Supervisor:

Dr. Laura Hollink (HCDA, CWI)

Dr. Adam S.Z. Belloum (UvA)

CWI

Centrum Wiskunde & Informatica

Outline

- Introduction
- Related work
- Methods & Evaluation
- Conclusion
- Future work

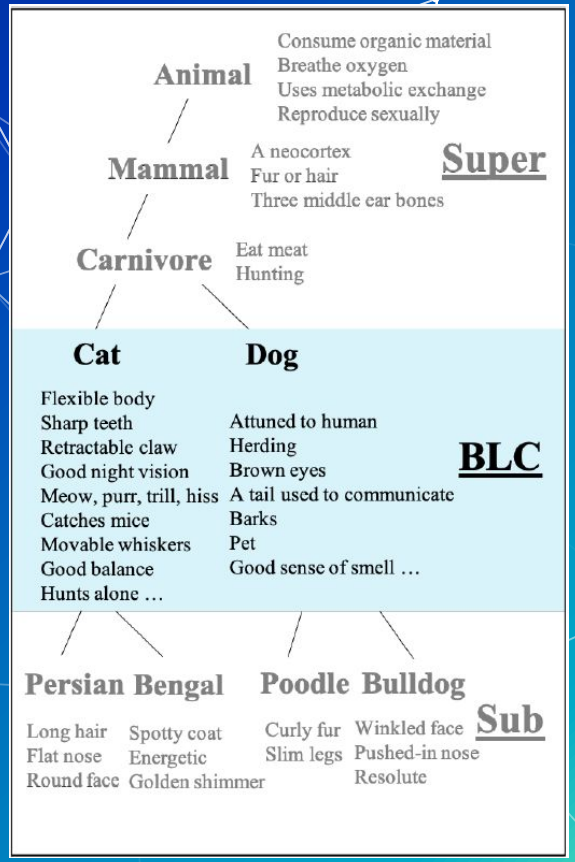
1.
Introduction to Basic Level

Definition, applications, and goals

Definition

Basic Level

- Basic-level categories
- By Rosch et al. in 1976
- Cognitive economy: react quicker and more accurately with less efforts
- Cue validity



Chen, Yiwen, and Simone Teufel. "Synthetic Textual Features for the Large-Scale Detection of Basic-level Categories in English and Mandarin." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.

Applications

Using Basic Level for categorization helps:

- Knowledge organization systems (KOS)
- Semantic Web
- Improve user interface

More benefits for:

- Natural language processing
- Computer vision

Goals - Research Questions

- **Predict basic level concepts from corpus characteristics and synthetic features**
  - Synthetic features: structural + frequency + semantics
- Corpus size - type(written/spoken) || audience(general/children)
- Semantic feature generation
- Detect in WordNet

2.
Related Research

Rule, Machine learning, and Context-aware basic level

Rule-Based Heuristics

- Mills et al.: Rule-based system with heuristics
  - Rules + Boundaries
  - Accuracy: 77.0%
  - 152 concepts

- Machine learning
- 840 concepts
- Accuracy: 89.9% (Cohen's kappa)

Chad Mills, Francis Bond, and Gina-Anne Levow. Automatic identification of basic-level categories. In Proceedings of the 9th Global Wordnet Conference, pages 298–305, 2018.

Machine Learning-Based Classification (1/3)

- HCDA: Predict in a concept hierarchy
  - Lexcical + structural + frequency features
  - Binary classification by Random Forest (best)
  - Hard for humans and the model
  - ❑ Semantic features: Word embeddings, Cues
  - ❑ Implicit + explict

Hollink Laura, Aysenur Bilgin, and Jacco van Ossenbruggen. "Predicting the basic level in a hierarchy of concepts." Research Conference on Metadata and Semantics Research. Springer, Cham, 2020.

Machine Learning-Based Classification (2/3)

- Henry: Find useful corpus properties
  - Larger corpus, more reliable result
  - Spoken corpus to children
  - Significance of type and size
  - Wilcoxon rank-sum test
  - Semantic features

Henry, N. “Learning the Basic Level from Text: Studying Different Corpus Characteristics in Predicting the Basic Level.” University of Amsterdam, 2021, https://scripties.uba.uva.nl/search?id=722721. Accessed 2 Dec. 2021.

Machine Learning-Based Classification (3/3)

- Chen & Teufel: Synthetic features at scale
  - Semantics: Wikipedia pages, Distributional Memory, and BART(best)
  - Synthetics: Cue validity, Basic level page rank, Semantics
  - SVM
  - Accuracy: 75%(English), 80%(Mandarin)

- ❑ Fine-tuned BART: cue generator
- ❑ Cue validity: to extract semantic features

Chen Yiwen, and Simone Teufel. "Synthetic Textual Features for the Large-Scale Detection of Basic-level Categories in English and Mandarin." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.

Context-Aware Basic Level in Folksonomies

- Chen et al.: Detect the basic level among context from folksonomies
  - Large-scale web resources
  - Implicit semantics
  - Contextual category utility
  - ❑ WordNet: a hierarchy of concepts
  - ❑ Lexical, structural, and frequency features

Wen-hao Chen, Yi Cai, Ho-fung Leung, and Qing Li. Context-aware basic level concepts detection in folksonomies. In International Conference on Web-Age Information Management, pages 632–643. 2010.

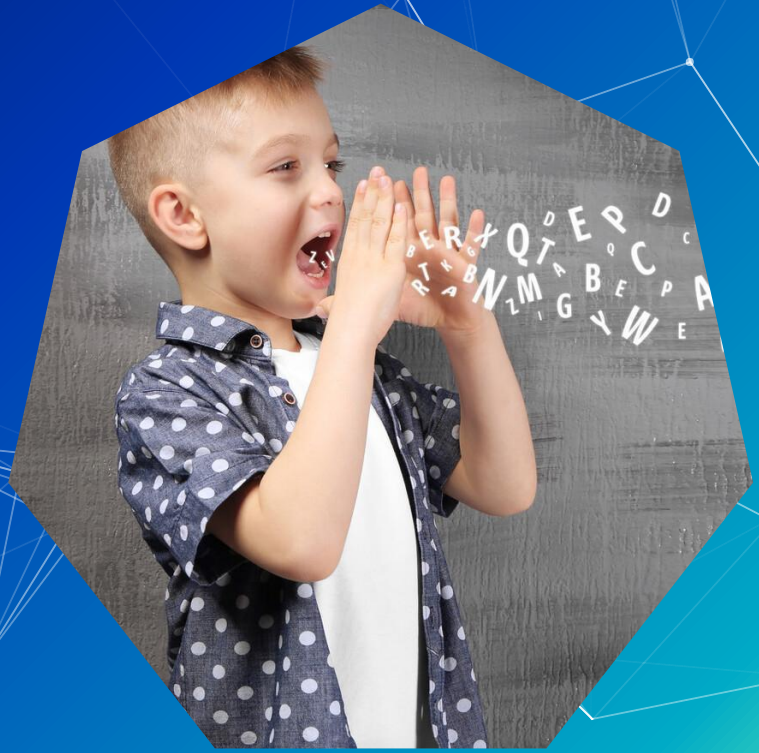
3.
Methods & Evaluation

Corpus characteristics, frequency, and semantic features
Results and interpretations

Sub-Research Question One

Improve from TYPE or AUDIENCE:

- Spoken better than written?
- Children better than general?



I. Corpus sampling - Setting

- **Corpora**

	Written	Spoken
General	**BNC**	**CABNC**
Children	**KBNC**	**CHILDES**

- **Size**

	1M	2.4M	5.7M	100M
BNC	✓	✓	✓	✓
CHILDES	✓	✓	✓	
CABNC	✓	✓		
KBNC	✓			

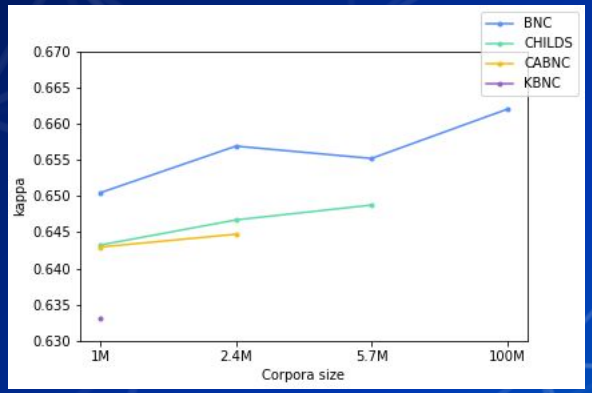
- **Random Forest Models**
  - Local: Same domain & 10-fold CV
  - Global: All domains & 10-fold CV
  - Transfer: The other domain

I. Corpus sampling

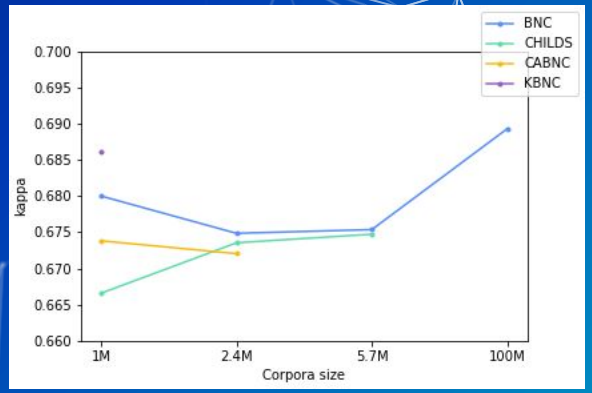
	Written	Spoken
General	BNC	CABNC
Children	KBNC	CHILDES

Metrics: Cohen's kappa (kappa)

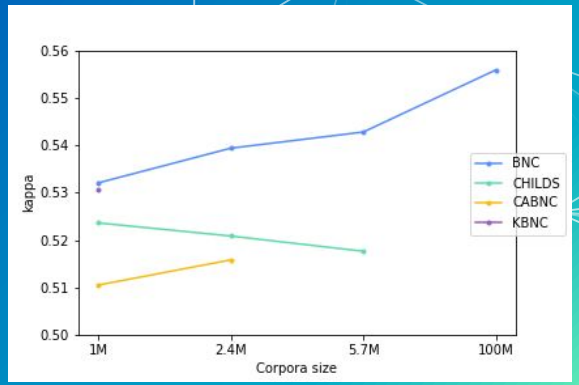
LocalModel:



GlobalModel:



TransferModel:



II. Wilcoxon Rank-sum Test on Size

- **Null hypothesis:**

  Results under the same model, the same corpus, but different two sizes are from the same distribution.

- **Alternative hypothesis**
  - Less
- **Bonferroni correction:** $\alpha' = \frac{0.05}{m}$
- **Variables**
  - Discourse type: written or spoken
  - Target audience: general or children
  - Combination

II. Wilcoxon Rank-sum Test on Size

- **TransferModel: (metric: kappa, m=9, α'=0.006)**

P-values (less)	Written	Spoken	General	Children
1M - 2.4M	**0.000**	0.129	**0.001**	0.99997
1M - 5.7M	**0.000**	0.246	**0.000**	1.0
1M - 100M	**0.000**		**0.000**	
2.4M - 5.7M	0.084	0.510	**0.000**	0.96235
2.4M - 100M	**0.000**		**0.000**	
5.7M - 100M	**0.000**		**0.000**	

- Limitation: KBNC(written children) only 1M
- Better with a larger corpus

III. Wilcoxon Rank-sum Test on Types

- **TransferModel (metric: kappa, m=3, α'=0.017)**

P-values (greater)	1M	2.4M	5.7M
Written - Spoken	**0.000**	**0.000**	**0.000**
General - Children	1.000	0.029	**0.000**

  - Written greater!
  - General greater?
  - Prove the limitation of KBNC 1M
- **Conclusions**:
  - Frequency from written general corpora
  - Large size

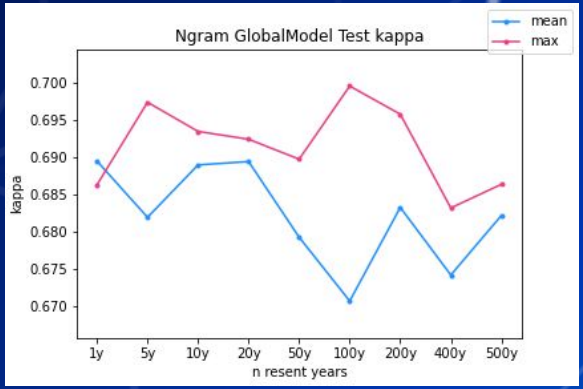
Larger Corpus: Google Books Ngram

- Written and general: printed
- Enormous size: 1500 - 2019
- Frequency search engine: Google Ngram Viewer
- Optimal Crawler +21% speedup

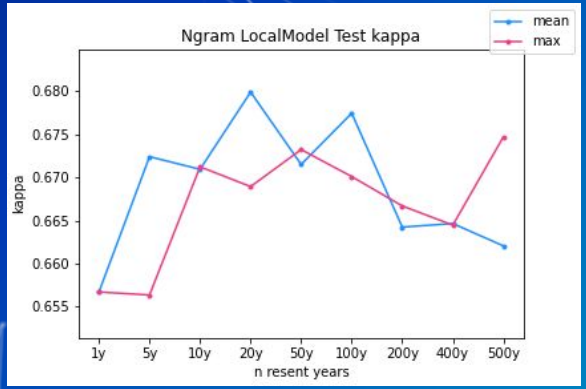
I. Frequency as a feature - Google Ngram

Mean and maximal frequencies in the recent 1 year, 5, 10, 20, 50, 100, 200, 400, and 500 years.

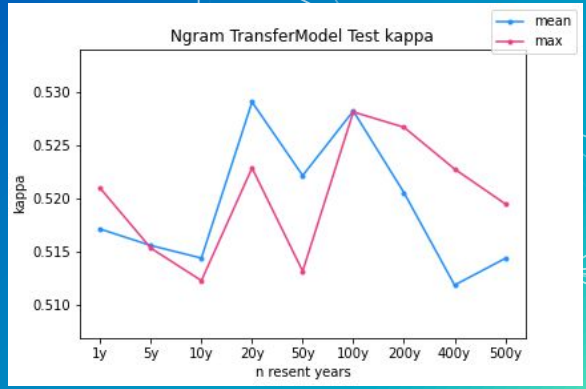
GlobalModel:



LocalModel:



TransferModel:



I. Frequency features - Google Ngram

- **Conclusions:**
  - No dependency on time periods
  - Mean or maximum
  - Needs of a feature selection
  - 50y and 100y for predictions in WordNet

II. Frequency Feature – Results

- **Best kappa**

Kappa	With Google Ngram	Baseline	Improvement	Growth rate
GlobalModel	0.714	0.673	0.041	+ 6.16%
LocalModel	0.712	0.640	0.072	+ 11.09%
TransferModel	0.590	0.521	0.069	+ 13.21%

- **Conclusions**
  - Much improvement due to **General audience** and **Written**
  - Better than BNC due to **SIZE**

Sub-Research Question Two

Improve from Semantics:

- Word embeddings
- Generated cues by BART

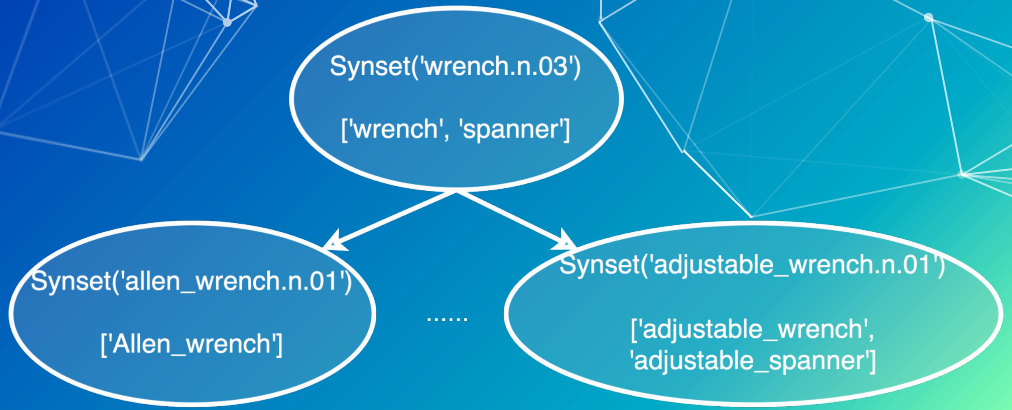
I. Word Embeddings – Word2vec

- **Two ways to represent semantics**
  - Vectors trained with SVM
  - Distance by vectors
- **Pre-trained W2V:** ConceptNet Numberbatch 19.08
- **From words to vectors**
  - 300 dimensions
  - Multiword concepts (e.g. ball-peen hammer): 63 removed

Robyn Speer, Joshua Chin, and Catherine Havasi (2017). "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge." In proceedings of AAAI 2017.

I. Word Embeddings - Distance

- **Semantics extraction**
  - Lemma Distance: Cosine similarity
  - Concept Distance: Hypernym lemmas ↔ Concept lemmas
  - Mean, minimum, maximum, and standard deviation



I. Word Embedding - Model Tests

- **Benchmark Model:** Random Forest with SMOTE
- **Results**

Kappa	With Semantic Distance	Baseline	Improvement	Growth rate
GlobalModel	0.713	0.673	0.040	+ 6.01%
LocalModel	0.648	0.640	0.008	+ 1.23%
TransferModel	0.531	0.521	0.010	+ 2.06%

II. Semantic Feature Generation (1/3)

- **Motivation**
  - English semantic features: properties, categories, attributes
  - Machine translation: Seq-to-Seq
  - BART: Pre-trained model from a denoising autoencoder
- **Indicator**
  - Textual → numerical
  - Cue validity

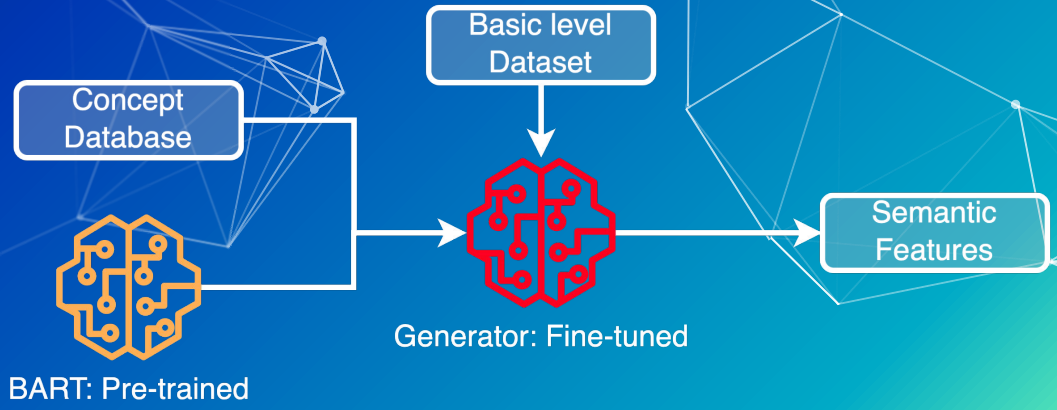
Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461 (2019).

II. Semantic Feature Generation (2/3)

- **Fine-tuned model**
  - Training data: a database of 4436 concepts

    E.g. 'abandon' : [desert, give up, leave, withdraw]

    (abandon, desert), (abandon, give up) ...
  - Method pipeline:
    - Tokenization
    - Fine-tuning
    - Generating



Buchanan, Erin M., Kathrene D. Valentine, and Nicholas P. Maxwell. "English semantic feature production norms: An extended database of 4436 concepts." Behavior Research Methods 51.4 (2019): 1849-1863.

II. Semantic Feature Generation (3/3)

- **Feature Transform**
  - **Statistics**
    - # cues
  - **Cue Validity**

$$CV(concept) = \sum_{cue \in semantics(concept)} P(BL|cue) = \sum_{cue} \frac{P(BL \wedge cue)}{P(cue)}$$

Rosch, Eleanor, et al. "Basic objects in natural categories." Cognitive psychology 8.3 (1976): 382-439.

II. Semantic Feature Generation : Model Tests

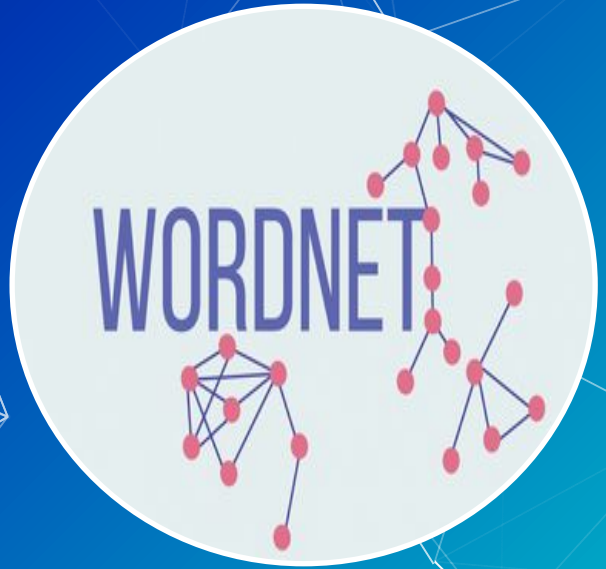
- **Benchmark Model:** Random Forest with SMOTE
- **Results**

Kappa	With Generated Cues	Baseline	Improvement	Growth rate
GlobalModel	0.881	0.673	0.208	+ 30.91%
LocalModel	0.898	0.640	0.258	+ 40.31%
TransferModel	0.705	0.521	0.184	+ 35.32%

Sub-research Question Three

Detect the basic level:

under Synset(“entity.n.01”) in WordNet



Pipeline: Extract the synthetic features

Structural features

- # direct hypernyms
- # total hyponyms
- normalized # part-whole
- normalized depth
- normalized gloss length
- shortest lemma length
- # lemmas
- maximal # polysemies

Frequency features

- Written
- To general audience
- Google Ngram
- 50-year
- 100-year

Semantic features

- Generate cues
- Calculate the cue validity

Random Forest with SMOTE

Comparison

- **Results** (74,374 concepts)

Prediction	Basic level	Agreement	Agreement
Hollink et al.	9,819	7,872	
Ours	9,085		8,588
Niamh Henry	15,401		

4.
Conclusion

Answer the research questions

Conclusion

- *SQ1: To what extent the **discourse type** and **target audience** of a corpus considering its **size** would affect the performance of predicting the basic level?*
  - Written, General corpora with a large size
- *SQ2: What new features concerning **semantics** can be generated to help improve the performance of predicting the basic level?*
  - Distance-based features from word embeddings
  - Cue validity from cues by the fine-tuned BART

Conclusion

- *SQ3: How much would corpus characteristics and synthetic features improve basic level detection in a large-scale hierarchy?*
- Good sensitivity
- No gold standard yet

5.
Future Work

Future Work

- Predicting
  - Multi-class
    - Basic level + superordinate + subordinate
  - Quantified regression
    - Score: probability
  - Generalization
    - Assemble learning

THANKS!

Basic Level Detection

Any questions?

Haochen Wang (CS, VU & UvA Joint Degree)

vrije Universiteit

CWI

Centrum Wiskunde & Informatica