Non-structured data construction

Renato Vukovic

Dialog Systems and Machine Learning
Institut für Informatik
Heinrich-Heine-Universität Düsseldorf

16.07.2021

Contents

Introduction

User: Hello, I am trying to find a place to stay that has **free wifi and 3 stars**. Do you have anything like that?
System: I have 5 results that match your criteria, do you have a certain **area of town** you'd like to stay in or a **price range** you're looking for?
User: Are any of them a **hotel**? I don't want a guesthouse.
System: Yes there are two hotels. One in the *centre* and one in the **south**. Any preference?
User: No preference, I need it for 2 nights starting from Monday for 3 people, and I need a reference number please.
System: OK, you are booked into the Gonville Hotel, located in the center of town on Gonville Place. The reference number is G6FAO56J.
User: That is great, I'm so excited. Thanks for you help. Bye.
System: Thank you for contacting the help desk, and have a great day.

Eric et al., 2019

Introduction

“Ontologies define the terms used to describe and represent an area of Knowledge.”

W3C (World Wide Web Consortium)

Motivation

- Most data is not readable for dialogue models.
- Structured data as used in ontologies is very limited, as they are mostly constructed by hand.
- Reducing human intervention in data structuring expands the amount of data which can be used.

Motivation

- Most data is not readable for dialogue models.
- Structured data as used in ontologies is very limited, as they are mostly constructed by hand.
- Reducing human intervention in data structuring expands the amount of data which can be used.

⇒ Including non-structured data in dialogue systems would make them richer in terms of what they could talk about.

Introduction

- Ärguably, the most challenging and time-consuming part of any dialogue data collection is the process of annotating dialogue acts."
- Multi-WOZ dataset has a size of around 10K dialogues.
  - 8 domains, 61 slots
  - 190 restaurants, 89 hotels with slots like "free internet", "food type", etc.
- Yelp review data contains more than 8M reviews for 160K businesses.
  - Likely has many more concepts than Multi-WOZ
  - Manual annotation most likely not feasible.

Budzianowski et al. 2018

https://www.yelp.com/dataset/

Ontology construction

- Manual construction
- Cooperative construction: Most tasks of the ontology construction supervised by human experts.
- (Semi-) Automatic construction: Reducing human intervention to a minimum.

Al-Aswadi et al., 2020

Yelp review example

"Great place to hang out after work: the prices are decent, and the ambience is fun. It’s a bit loud, but very lively. The staff is friendly, and the food is good. They have a good selection of drinks."

https://www.yelp.com/dataset/

Yelp review example

"Great place to hang out after work: the prices are decent, and the ambience is fun. It's a bit loud, but very lively. The staff is friendly, and the food is good. They have a good selection of drinks."

https://www.yelp.com/dataset/

Procedure

- Goal: Develop a process for unsupervised ontology building or knowledge extraction for use in dialogue models without additional training.

Budzianowski et al. 2018

Procedure

- Goal: Develop a process for unsupervised ontology building or knowledge extraction for use in dialogue models without additional training.
  1. Rebuild the Multi-WOZ ontology as automatically as possible.

Budzianowski et al. 2018

Procedure

- Goal: Develop a process for unsupervised ontology building or knowledge extraction for use in dialogue models without additional training.
  1. Rebuild the Multi-WOZ ontology as automatically as possible.
  2. Use this process to build a Multi-WOZ like ontology from unlabeled Yelp review data.

Budzianowski et al. 2018

Procedure

- Goal: Develop a process for unsupervised ontology building or knowledge extraction for use in dialogue models without additional training.
  1. Rebuild the Multi-WOZ ontology as automatically as possible.
  2. Use this process to build a Multi-WOZ like ontology from unlabeled Yelp review data.
  3. Add new slots, e.g. politeness of staff, ambience, etc.

Budzianowski et al. 2018

Procedure

- Goal: Develop a process for unsupervised ontology building or knowledge extraction for use in dialogue models without additional training.
  1. Rebuild the Multi-WOZ ontology as automatically as possible.
  2. Use this process to build a Multi-WOZ like ontology from unlabeled Yelp review data.
  3. Add new slots, e.g. politeness of staff, ambience, etc.
  4. Find new Concepts possibly in an unsupervised manner.

Budzianowski et al. 2018

Ontology construction

1. Extraction
2. Concept discovery
3. Concept refining
4. Finding relations between concepts

Al-Aswadi et al., 2020

Ontology construction

1. **Extraction**
2. Concept discovery
3. Concept refining
4. Finding relations between concepts

Term extraction

- Assumption: Correlation between word frequency and importance

Henderson, et al., 2020

Term extraction

- Assumption: Correlation between word frequency and importance
  - $\Rightarrow$ Extract keyphrases based on frequency.

Henderson, et al., 2020

Term extraction

- Assumption: Correlation between word frequency and importance
  $\Rightarrow$ Extract keyphrases based on frequency.
- Cover all the important slots, values in this step.

Henderson, et al., 2020

Term extraction

- Assumption: Correlation between word frequency and importance
  $\Rightarrow$ Extract keyphrases based on frequency.

- Cover all the important slots, values in this step.
- Filter unnecessary words which do not refer to concepts later.

Henderson, et al., 2020

Term extraction

- Extract the highest scoring keyphrases from the data using the scoring function: $s(k) = (\prod_{i=1}^{n} freq(w_i))^{1/n^{\alpha}}$ for words $(w_1, w_2, ..., w_n)$ from keyphrase k
- Add entities, which are recognised by a named entity recogniser(NER).

Henderson, et al., 2020

Term extraction examples

- ’What is the address, phone number, and **price range** of the grafton hotel restaurant?’
- ’I want to get there **by 19:45** at the latest.’
- ’I am **departing from birmingham** new street.’
- "No, I just need to **make sure** it’s cheap. oh, and I need parking."

Keyphrase tuning

$$s(k) = \left(\prod_{i=1}^{n} freq(w_i)\right)^{1/n^{\alpha}}$$

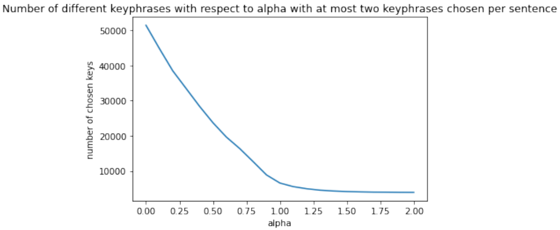
Keyphrase tuning

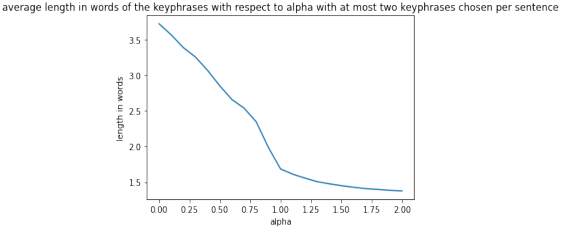
$s(k) = \left(\prod_{i=1}^{n} freq(w_i)\right)^{1/n^{\alpha}}$

- Change α

Changing α

Heinrich Heine
Universität Düsseldorf





Keyphrase tuning

Heinrich Heine
Universität Düsseldorf

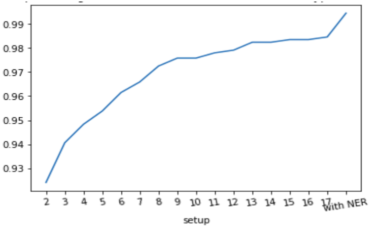
$s(k) = (\prod_{i=1}^{n} freq(w_i))^{1/n^{\alpha}}$

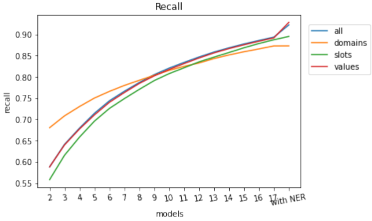
- Change α
- Number of keyphrases taken per turn

Number of keyphrases per turn + NER

Heinrich Heine
Universität Düsseldorf

Percentage of values covered wrt. the number of keyphrases per turn



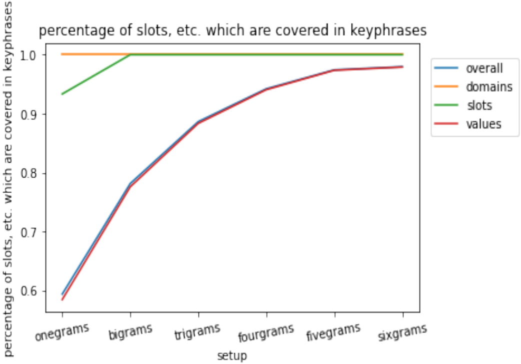


Keyphrase tuning

$s(k) = (\prod_{i=1}^{n} freq(w_i))^{1/n^{\alpha}}$

- Change α
- Number of keyphrases taken per turn
- Keyphraselength

Keyphrase length



Ontology construction

1. Extraction
2. **Concept discovery**
3. Concept refining
4. Finding relations between concepts

Concept discovery

- Train the model on predicting whether the extracted keyphrases are present.

Concept discovery

- Train the model on predicting whether the extracted keyphrases are present.

  $\Rightarrow$ Similar concepts are close in the embedding space.

Concept discovery

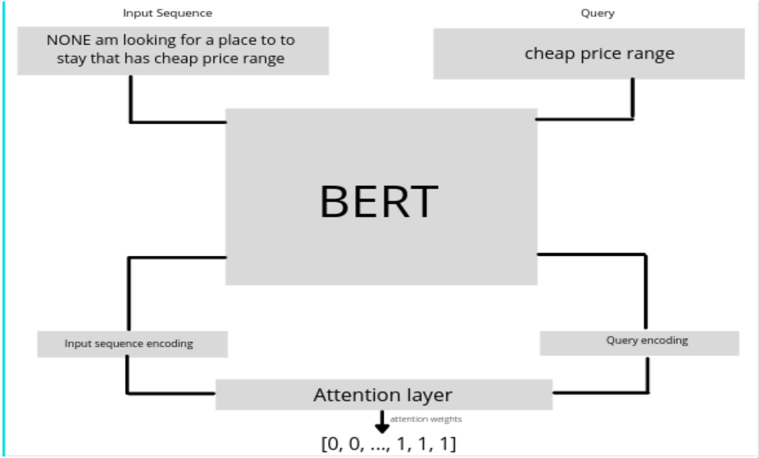
- Train the model on predicting whether the extracted keyphrases are present.
  $\Rightarrow$ Similar concepts are close in the embedding space.
- Model tags the queried keyphrase, if it is present.

Concept discovery

Heinrich Heine Universität Düsseldorf

- Train the model on predicting whether the extracted keyphrases are present.
  $\Rightarrow$ Similar concepts are close in the embedding space.
- Model tags the queried keyphrase, if it is present.
- Attends to "NONE"token if the query is not present in the input sequence.

Pretraining



Heck, experimental TripPy version

Concept discovery

examples after keyphrase training

- Model tags "No, I just need to make sure it’s **cheap**. oh, and I need parking."with query "price"
- With query "free": "No, I just **need** to make sure it’s **cheap. oh**, and I **need parking**."

Concept discovery

- Query slot-values from the Multi-WOZ dataset with the pretrained model.
- Model able to classify a queried slot-value as present in input sequence with 82.4% accuracy Multi-WOZ 2.2 test set.

Zang et al. 2020

Ontology construction

1. Extraction
2. Concept discovery
3. **Concept refining**
4. Finding relations between concepts

Concept refining

- Train the model on the slot-value pairs from the Multi-WOZ dataset, so that it is able to get these concepts and tag them.
- Filter the unnecessary parts of the keyphrases this way, i.e. increase the precision.

Ontology construction

Heinrich Heine
Universität Düsseldorf

1. Extraction
2. Concept discovery refining
3. Concept refining
4. **Finding relations between concepts**

Extract concepts from unlabeled data

Heinrich Heine Universität Düsseldorf

- Apply the model trained on tagging slot-value pairs in Multi-WOZ on Yelp data.

Extract concepts from unlabeled data

Heinrich Heine
Universität Düsseldorf

- Apply the model trained on tagging slot-value pairs in Multi-WOZ on Yelp data.
- Query the known slots and tag their corresponding values in the reviews (e.g. price-decent).

Extract concepts from unlabeled data

- Apply the model trained on tagging slot-value pairs in Multi-WOZ on Yelp data.
- Query the known slots and tag their corresponding values in the reviews (e.g. price-decent).
- Introduce new slots found in the new dataset (e.g. staff-friendly).

Conclusion

Heinrich Heine Universität Düsseldorf

- Structuring huge amounts of data is a difficult challenge of current research.
- Automatic processes to build knowledge bases could make models more versatile, if they can be used sufficiently.
- It is unclear whether building a knowledge base in a fully automatic way is possible.
- Using known ontologies to develop a process seems promising.

Heinrich Heine
Universität Düsseldorf

Thanks for your attention!

References

- Matthew Henderson and Ivan Vulić, PolyAI Ltd, London, UK, Language Technology Lab, University of Cambridge, UK ConVEx: Data-Efficient and Few-Shot Slot Labeling, 2020
- Budzianowski, Paweł and Wen, Tsung-Hsien and Tseng, Bo-Hsiang and Casanueva, Iñigo and Ultes Stefan and Ramadan Osman and Gašić, Milica, MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2018
- yelp academic dataset `https://www.yelp.com/dataset/`

References

- MultiWOZ 2.1: Multi-Domain Dialogue State Corrections and State Tracking Baselines, Eric, Mihail and Goel, Rahul and Paul, Shachi and Sethi, Abhishek and Agarwal, Sanchit and Gao, Shuyag and Hakkani-Tur, Dilek, 2019
- Automatic ontology construction from text: a review from shallow to deep learning trend, Fatima N. Al-Aswadi, Huah Yong Chan, Keng Hoon Gan, 2020
- MultiWOZ 2.2: A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines, Zang, Xiaoxue and Rastogi, Abhinav and Sunkara, Srinivas and Gupta, Raghav and Zhang, Jianguo and Chen, Jindong, Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020