Community Question Answering (cQA)

Outline

- Motivations
  - Limitations of Factoid QA
- Tasks
  - Question retrieval
  - Question duplicate detection
  - Question answering
- Applications
  - Automatic ticket answering
  - Semi-automatic FAQ building
- Datasets
  - Quora, SemEval2016, Yahoo!, CIKM 2018 Analytic Cup
- SOTA models
  - Kernel-based models
  - Neural Network models

Motivations

Factoid QA

- Learning systems that automatically answer questions posed by users in natural language.
- It deals mainly with **factoid** questions
  - questions that require a name as answer.
    - **Q:** Who is the president of US? **A:** *Donald Trump*.
- Answers from high-quality corpora
  - e.g. Wikipedia
- Useful?

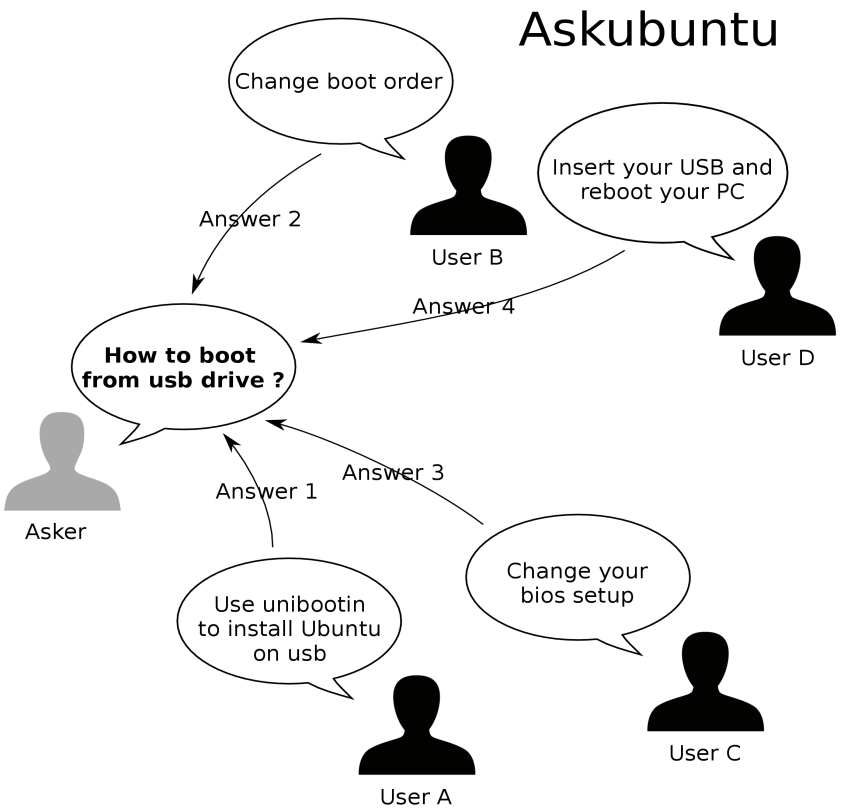
Limitations of Factoid QA

- 50% of questions asked on search engines, e.g. Google, are non-factoid (Berger et al., 2000)
  - How do I fix my car wheel?
  - What is the best restaurant in Trento?
- Answers cannot be found on standard knowledge repositories
  - e.g. Wikipedia
- Other sources of information are needed.

Social Media

- Computer technologies for creating and sharing of information.
- Contains answers to many daily questions asked by users
  - Quora, Yahoo! Answers, StackOverflow, AskUbuntu, etc..

Social Media



Automatic community Question Answering (cQA)

- Building systems that can automatically answer questions asked by users on **social media** websites.
- Use questions and answers generated by users, i.e. UGC, as training material.
    - Q&A websites, forums, etc..

Tasks

Question retrieval

- Ranking a set of retrieved questions in terms of their similarity with the original question.

Question retrieval

Original Question	Related Questions	Relevancy
Q1: Can i extend my family visit visa after 6 month??	**Q1_R1:** How to Convert Tourist visa to Family visit visa?	0
	Q1_R2: Maximum period of a Visit Visa?	0
	Q1_R3: Re-apply family visit visa?	**1**
	Q1_R4: How to extend the visa for my sister in law?	**1**
	Q1_R5: Family visit visa extension in Immigration?	**1**
	Q1_R6: New Family Visit Visa Extension	**1**

Question duplicate detection

- Detecting if two questions are semantically similar.

Question duplicate detection

Question1	Question2	is_duplicate
1: How can I be a good geologist?	**2**: What should I do to be a great geologist?	**1**
3: What can make Physics easy to learn?	**4:** How can you make physics easy to learn?	**1**
5: What is web application?	**6:** What is the web application framework?	0
7: Why do rockets look white?	**8**: Why are rockets and boosters painted white?	**1**
9: What is best way to make money online?	**10**: What is best way to ask for money online?	0

Question Answering

- Given a question and a set of candidate answers, reranking the answers according to their relevance wrt the question.

Question Answering

Question	Candidate Answers	Relevancy
Q1: Is it a Good Offer 800 QR?	**A1**: Depending on what is your work	0
	A2: it depends. Where u hired?	0
	A3: What grade are u being offered?	0
	A4: As a single… yes	**1**
	A5: The salary is good	**1**

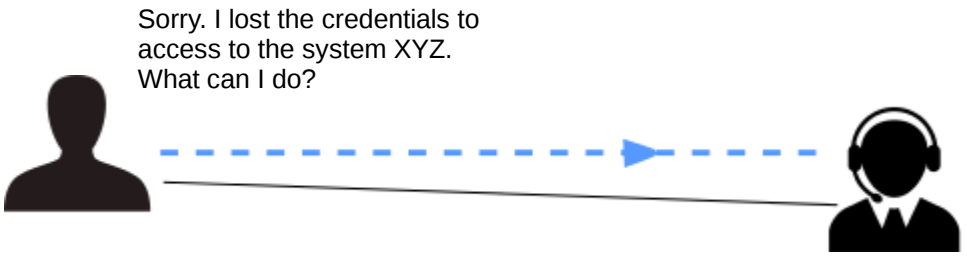
Applications

Automatic Ticket Answering

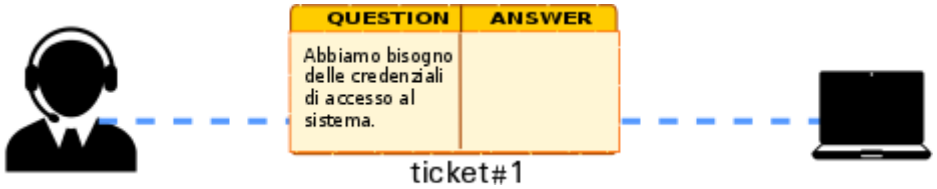
- Many companies run help desk (HD) offices for supporting customers.
- Help Desk offices uses ticketing systems to keep track of questions (and answers) asked by costumers of a company selling goods or services.
- **Task:** Automatically answering new ticket questions asked by users and stored in HD systems.

Ticket Answering

1. Users of a company call operators working in an HD service to ask questions or solicit actions.



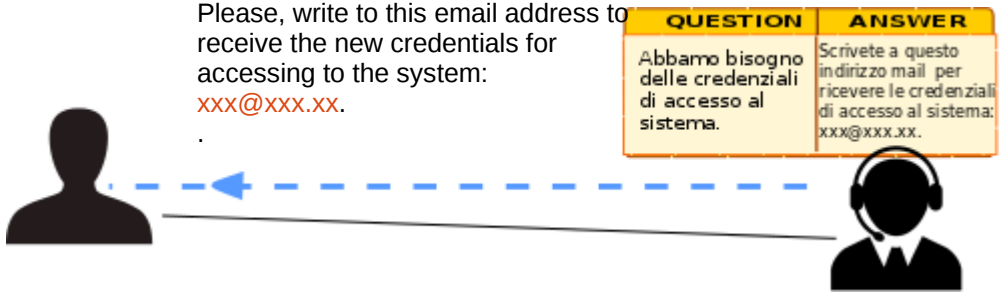
2. The operator takes in charge the user request and stores it in a ticket.



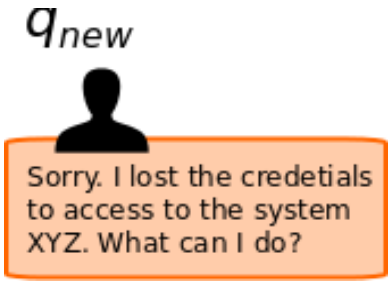
3. Operators search for answers among different sources.

- e.g. past tickets, FAQ, forums, user manuals and domain knowledge.

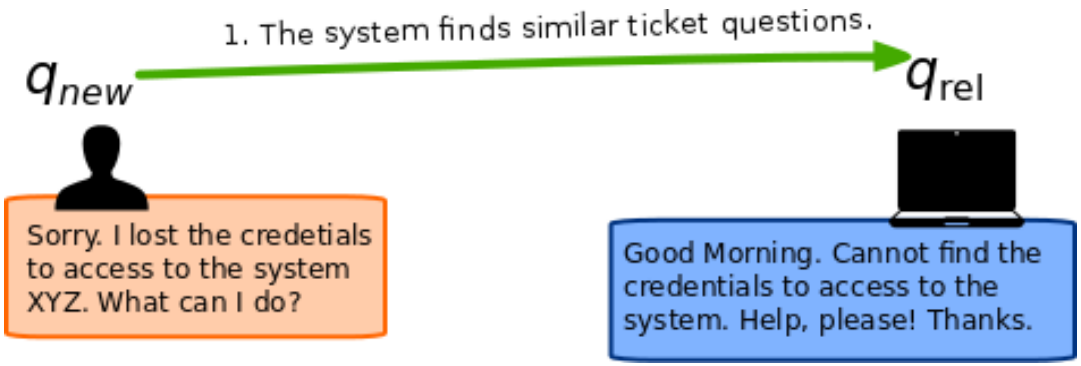
4. The answer is stored in the ticket and returned to the client.



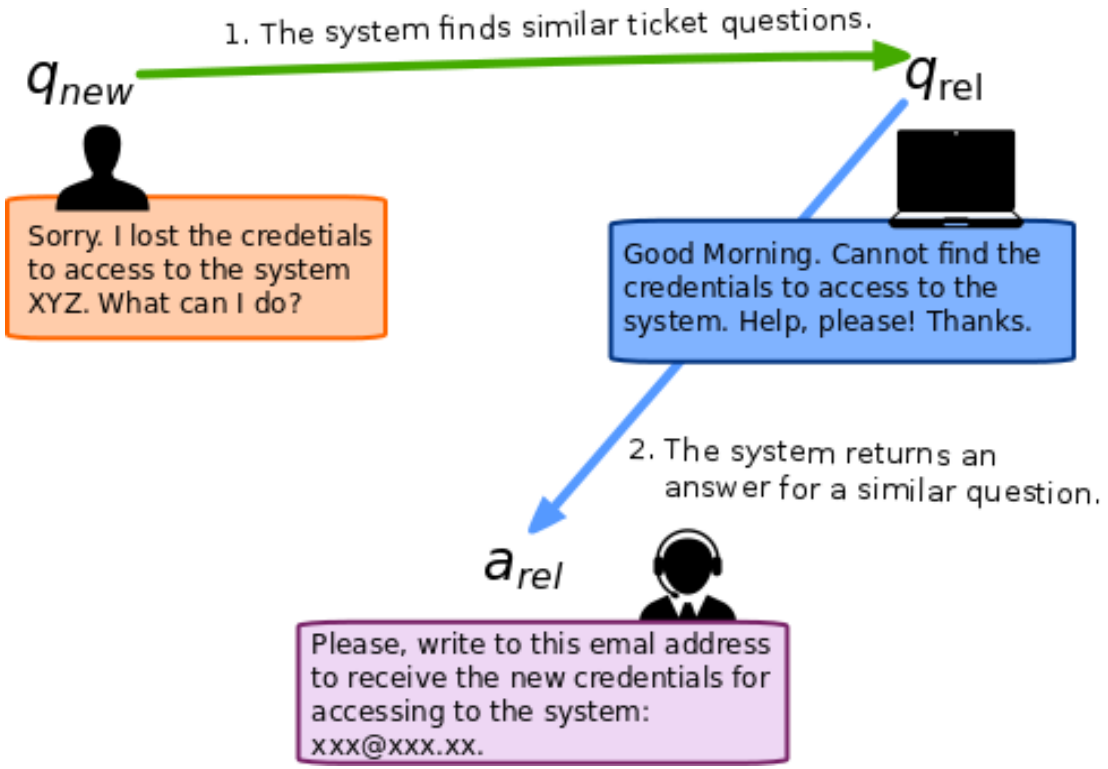
Automatic Ticket Answering



Automatic Ticket Answering



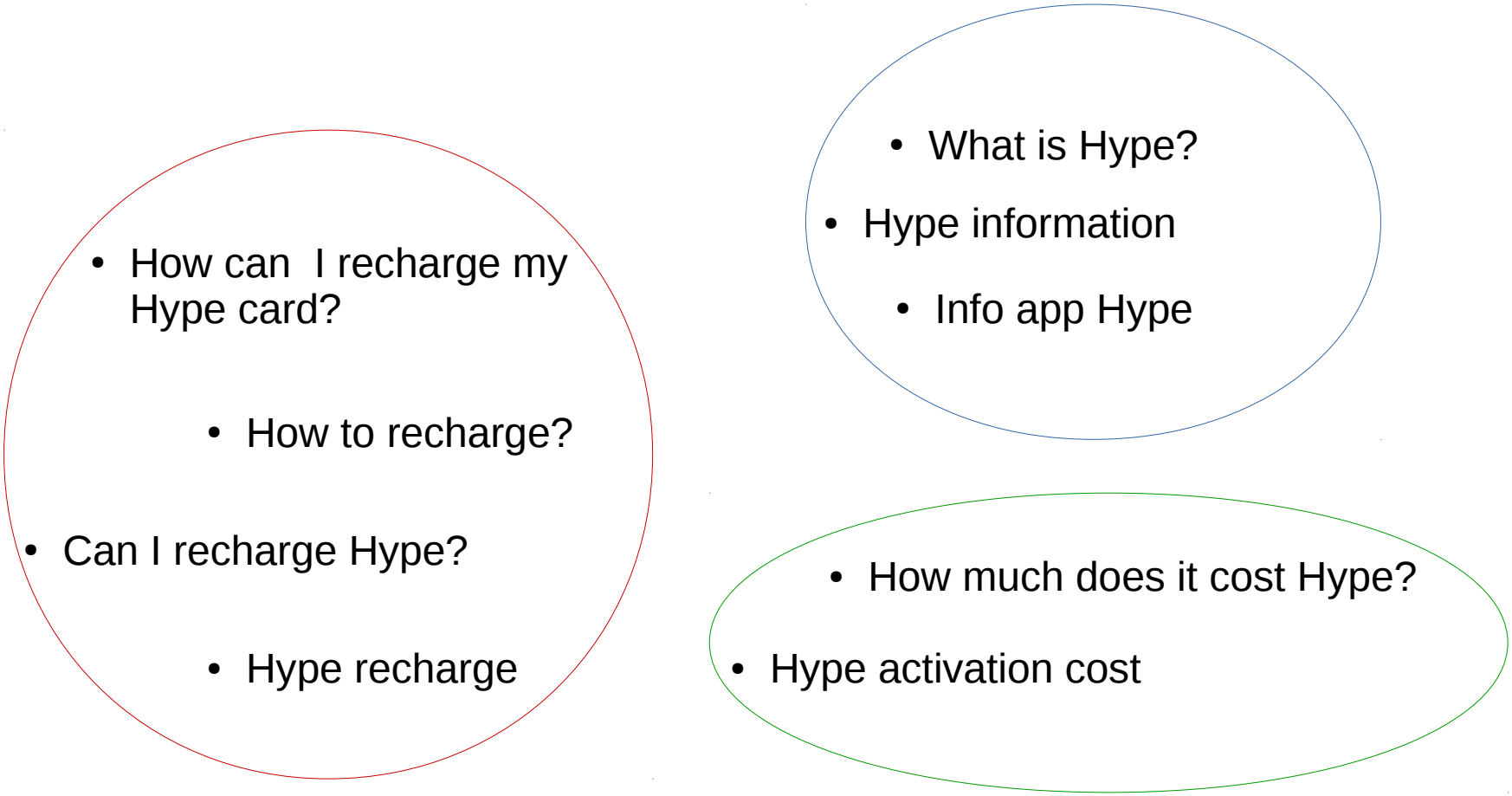
Automatic Ticket Answering



(Semi-)automatic FAQ building

- Grouping questions having the same intent for automatically building a list of FAQs.
- **Why:** Useful if you don't know in advance what are the kind of questions users are going to ask more frequently.

(Semi-)automatic FAQ building



Datasets

Quora

- **Task:** question duplicate detection
  - Pairs marked as 1 (duplicate) or 0 (not-duplicate)
- **Data:**
  - 404,348 question pairs from *Quora*:
    - 384,348 train set
    - 10,000 Dev. set
    - 10,000 test set
- **Pro:**
  - Large dataset

Cons:

- not annotated by human experts

- **Evaluation Metrics:** Accuracy, logloss

SemEval2016 CQA-QL

- **Task B:** question retrieval
- **Data:**
  - 3,469 question pairs from Qatar Living forum
    - 2669 train set
    - 500 dev. set
    - 700 test set
  - Each question composed of a subject and a body
- **Pro:**
  - Annotated by human experts

Cons:

– small dev. and test set

- **Evaluation Metrics:** MAP, MRR, P@1

Yahoo! QA

- **Task:** Question Answering
- **Data**:
  - 253,440 QA pairs in train set
  - 31,680 pairs in dev. Set
  - 31,680 pairs in test set
- **Pro:**
  - Large dataset

Cons:

– not annotated by experts

- **Evaluation Metrics** MAP, P@1

CIKM 2018 Analytic Cup
<Ongoing challenge>

- **Task:** Cross-lingual question duplicate detection
- **Data:**
  - 21,400 pairs in train set **(both English and Spanish)**
  - 5,000 pairs in test set (only Spanish)
- **Pro:**
  - Medium size
  - Annotated by human experts
- **Evaluation Metrics:** logloss

State of the art for Short Text Matching

SOTA

- Tree Kernel (Da SanMartino et al., 2016)
- CNN (Severyn et al., 2015)
- RNN/Tree LSTM (Tai et al., 2015)
- Attentive Networks (Parikh et al., 2016)
- Siamese Networks (Nicosia et al., 2016)

Tree Kernel (Da San Martino et al., 2016)

- Learn a function that measures the degree of syntactic similarity between two questions/phrases.

Tree Kernel (Da San Martino et al., 2016)

- Learn a function that measures the degree of syntactic similarity between two questions/phrases.

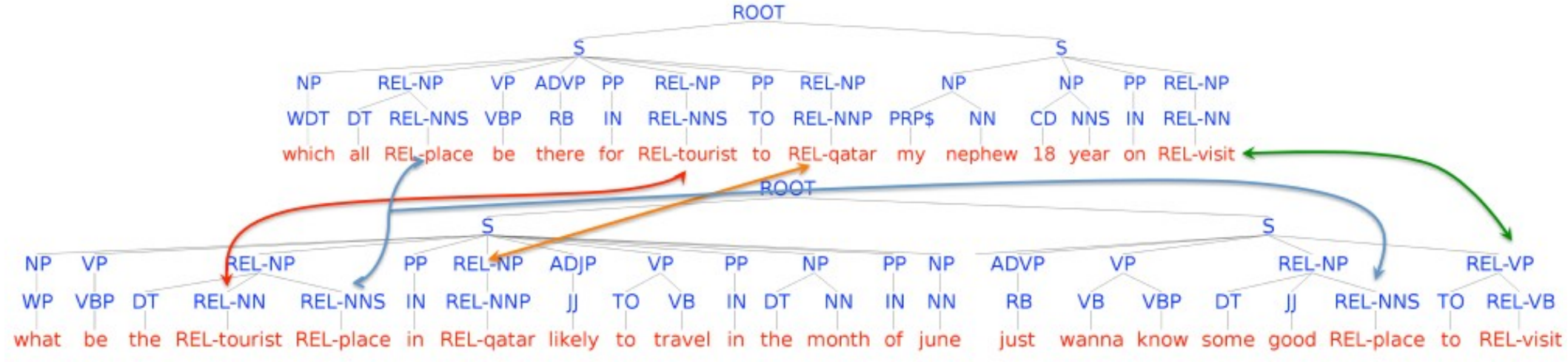
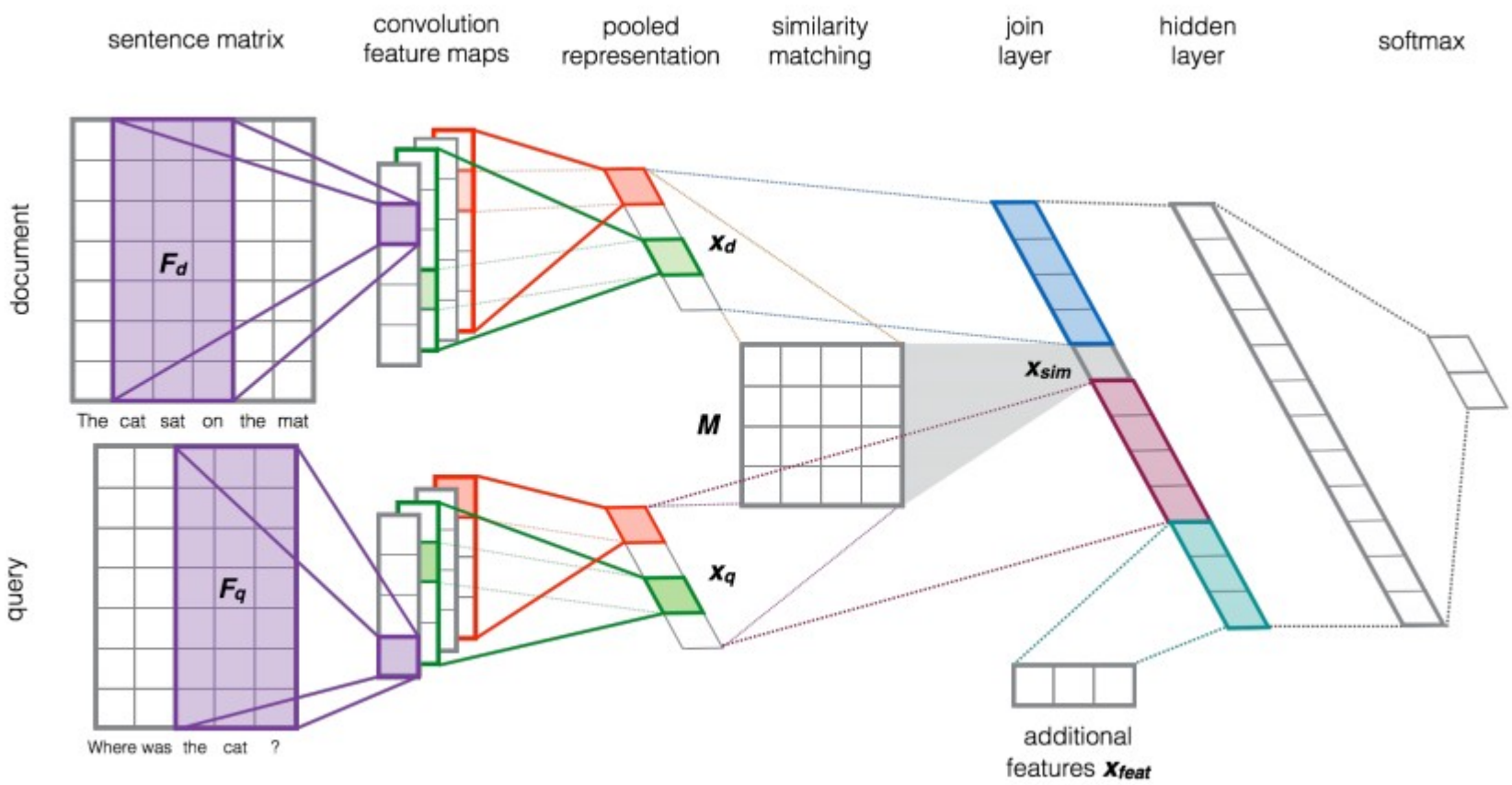


Figure 1: Our representation based on syntactic trees for the q_o–q_s pairs enriched with REL links.

CNN (Severyn and Moschitti, 2015)

- Use distributional sentence models to encode question/answer pairs in a vector.
  - Apply a set of convolution operations.
- Learn a function that measures the relevancy of answer wrt a question.
- Originally proposed for Question Answering.

CNN (Severyn and Moschitti, 2015)

- 

Tree LSTM (Tai et al., 2015)

- Recurrent Neural Network (RNN)
- composes its state from input vector and the hidden state of many child units.
- Tree-structured topology
  - Better at representing sentence meaning than sequential LSTM
- Proposed for semantic relatedness and sentiment classification

Tree LSTM (Tai et al., 2015)

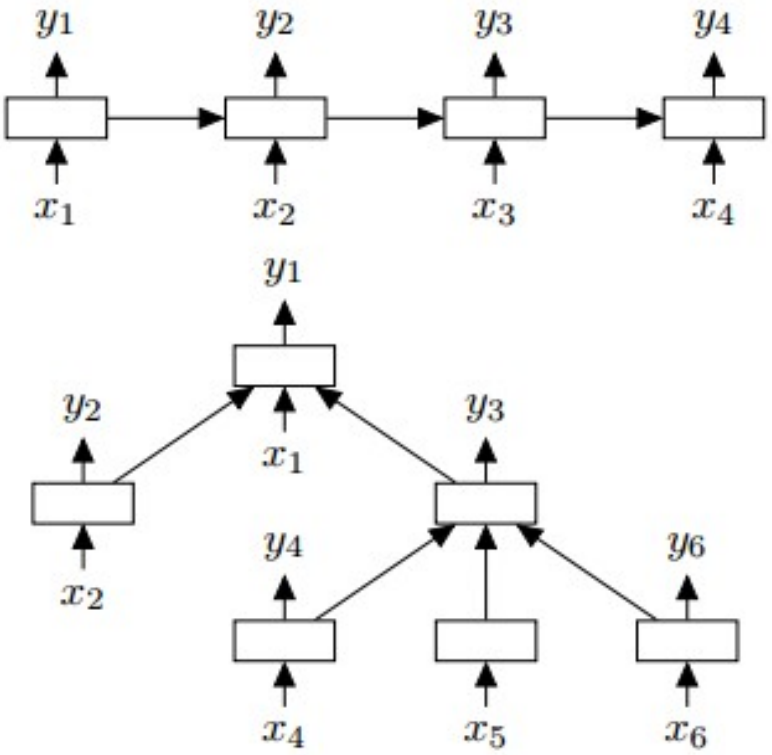


Figure 1: **Top:** A chain-structured LSTM network. **Bottom:** A tree-structured LSTM network with arbitrary branching factor.

Attentive Networks (Parikh et al., 2016)

- Compute soft-alignment matrix
- Decompose task into subproblems:
  - **Attend:** soft-align word groups in two sentences
  - **Compare:** compare aligned subphrases
  - **Aggregate:** aggregate results of previos step and output similarity.

Attentive Networks (Parikh et al., 2016)

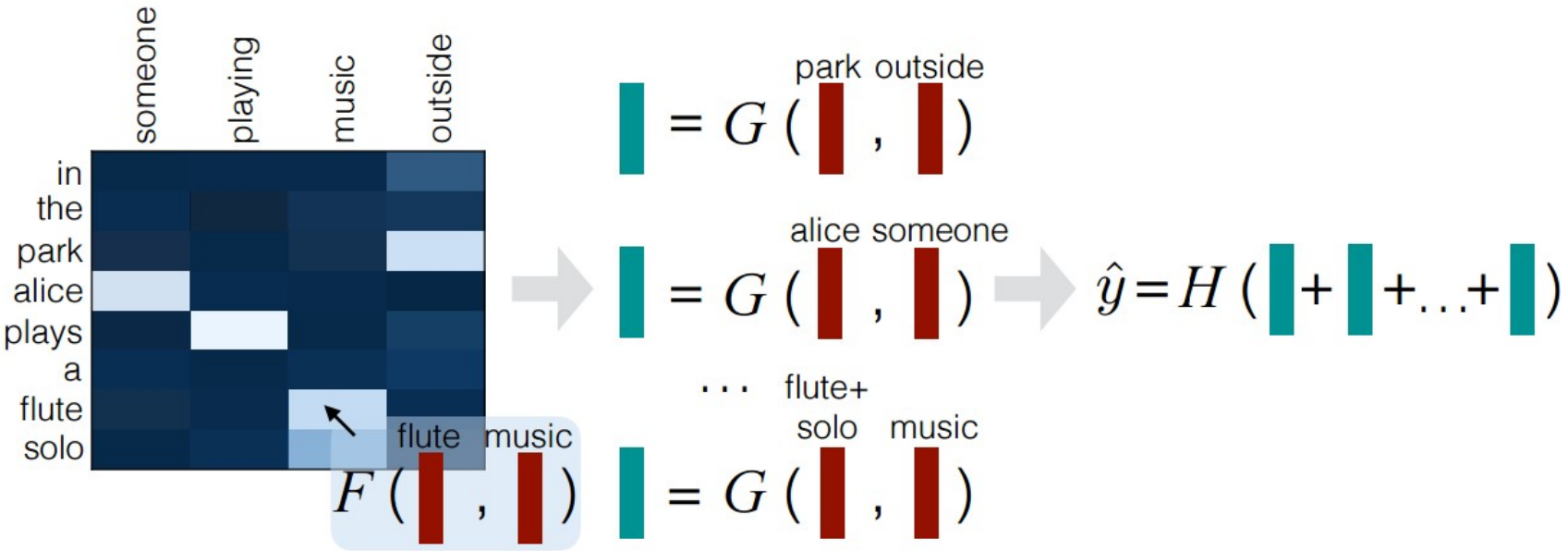
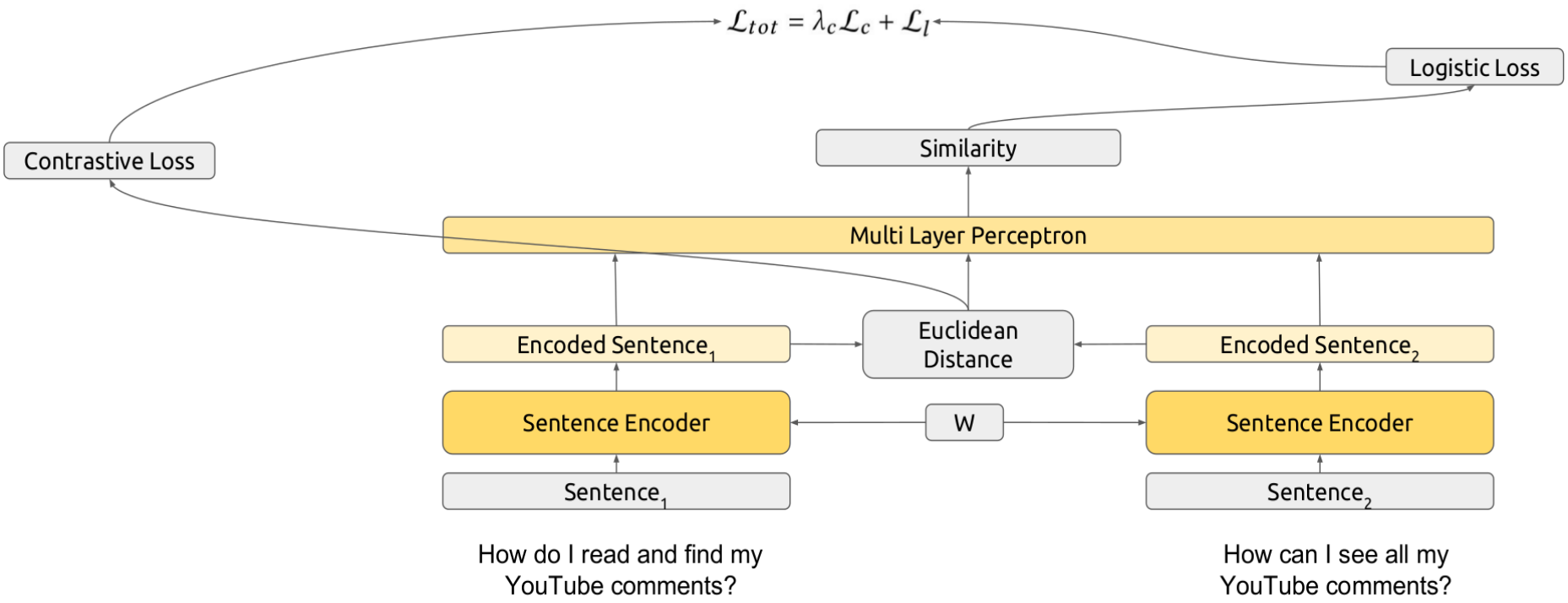


Figure 1: Pictoral overview of the approach, showing the *Attend* (left), *Compare* (center) and *Aggregate* (right) steps.

Siamese NN (Nicosia and Moschitti, 2017)

- Map two questions into the same geometric space
- Optimize contrastive loss by penalizing:
  - different question pairs judged having high cosine similarity by the NN.
  - similar question pairs judged having low cosim by the NN.

Siamese NN (Nicosia and Moschitti, 2017)



References

- Berger, A., Caruana, R., Cohn, D., Freitag, D., and Mittal, V. (2000). Bridging the lexical chasm: statistical approaches to answer-finding. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 192–199. ACM
- Da San Martino, Giovanni, et al. "Learning to re-rank questions in community question answering using advanced features." Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016.
- Severyn, Aliaksei, and Alessandro Moschitti. "Learning to rank short text pairs with convolutional deep neural networks." Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015.
- Tai, Kai Sheng, Richard Socher, and Christopher D. Manning. "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks."
- Parikh, Ankur, et al. "A Decomposable Attention Model for Natural Language Inference." Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016.
- Nicosia, Massimo, and Alessandro Moschitti. "Accurate Sentence Matching with Hybrid Siamese Networks." Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 2017.