ITCS 4111/5111: Intro to Natural Language Processing

Coreference Resolution

Razvan C. Bunescu

Department of Computer Science @ CCI

razvan.bunescu@uncc.edu

Coreference Resolution: Wikipedia text

- Determine which noun phrases (**mentions**) refer to the same discourse **entity**.

Originally from Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree.

Originally from Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree.

CoreferenceResolution: Mathematical Statements

- **Theorem**: $\sqrt{2}$ is an irrational number.
- **Proof**: Suppose that $\sqrt{2}$ were a rational number, so by definition $\sqrt{2} = a / b$ where a and b are non-zero integers with no common factor. Thus, $b\sqrt{2} = a$. Squaring both sides yields $2b^2 = a^2$. Since 2 divides the left hand side, 2 must also divide the right hand side (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write $a = 2c$, where c is also an integer. Substitution into the original equation yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that $\sqrt{2}$ is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: $\sqrt{2}$ is an irrational number.
- **Proof**: Suppose that $\sqrt{2}$ were a rational number, so by definition $\sqrt{2} = a / b$ where a and b are non-zero integers with no common factor. Thus, $b\sqrt{2} = a$. Squaring both sides yields $2b^2 = a^2$. Since 2 divides **the left hand side**, 2 must also divide the right hand side (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write $a = 2c$, where c is also an integer. Substitution into the original equation yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that $\sqrt{2}$ is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: √2 is an irrational number.
- **Proof**: Suppose that √2 were a rational number, so by definition √2 = a / b where a and b are non-zero integers with no common factor. Thus, b√2 = a. Squaring both sides yields **$2b^2$** = a^2. Since 2 divides **the left hand side**, 2 must also divide the right hand side (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write a = 2c, where c is also an integer. Substitution into the original equation yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that √2 is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: $\sqrt{2}$ is an irrational number.
- **Proof**: Suppose that $\sqrt{2}$ were a rational number, so by definition $\sqrt{2} = a / b$ where a and b are non-zero integers with no common factor. Thus, $b\sqrt{2} = a$. Squaring both sides yields $\mathbf{2b^2} = a^2$. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write $a = 2c$, where c is also an integer. Substitution into the original equation yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that $\sqrt{2}$ is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: $\sqrt{2}$ is an irrational number.
- **Proof**: Suppose that $\sqrt{2}$ were a rational number, so by definition $\sqrt{2} = a / b$ where a and b are non-zero integers with no common factor. Thus, $b\sqrt{2} = a$. Squaring both sides yields $\mathbf{2b^2} = \mathbf{a^2}$. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write $a = 2c$, where c is also an integer. Substitution into the original equation yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that $\sqrt{2}$ is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: √2 is an irrational number.
- **Proof**: Suppose that √2 were a rational number, so by definition √2 = a / b where a and b are non-zero integers with no common factor. Thus, b√2 = a. Squaring both sides yields **$2b^2$** = **a^2**. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write a = 2c, where c is also an integer. Substitution into **the original equation** yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that √2 is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: √2 is an irrational number.
- **Proof**: Suppose that √2 were a rational number, so by definition √2 = a / b where a and b are non-zero integers with no common factor. Thus, b√2 = a. Squaring both sides yields **$2b^2 = a^2$**. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write a = 2c, where c is also an integer. Substitution into **the original equation** yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that √2 is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: √2 is an irrational number.
- **Proof**: Suppose that √2 were a rational number, so by definition √2 = a / b where a and b are non-zero integers with no common factor. Thus, b√2 = a. Squaring both sides yields **$2b^2 = a^2$**. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write a = 2c, where c is also an integer. Substitution into **the original equation** yields $2b^2 = (2c)^2 = 4c^2$. Dividing **both sides** by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that √2 is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: √2 is an irrational number.
- **Proof**: Suppose that √2 were a rational number, so by definition √2 = a / b where a and b are non-zero integers with no common factor. Thus, b√2 = a. Squaring both sides yields **$2b^2 = a^2$**. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write a = 2c, where c is also an integer. Substitution into **the original equation** yields **$2b^2$** = $(2c)^2$ = **$4c^2$**. Dividing **both sides** by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that √2 is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: √2 is an irrational number.
- **Proof**: Suppose that √2 were a rational number, so by definition √2 = a / b where a and b are non-zero integers with no common factor. Thus, b√2 = a. Squaring **both sides** yields $\mathbf{2b^2 = a^2}$. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as **they** are equal and **both** integers). So a^2 is even, which implies that a must also be even. So we can write a = 2c, where c is also an integer. Substitution into **the original equation** yields $\mathbf{2b^2} = (2c)^2 = \mathbf{4c^2}$. Dividing **both sides** by 2 yields $b^2 = 2c^2$. But then, by **the same argument as before**, 2 divides b^2, so b must be even. However, if a and b are **both** even, **they** share a factor, namely 2. **This** contradicts **our assumption**, so we are forced to conclude that √2 is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Application: Social Networks from Text

- Input = a large text, such as a novel or a play, for example *Romeo and Juliet*, which contains a number of characters, e.g. Romeo Montague, Juliet Capulet, Mercutio, Benvolio.
- Output = a social network extracted from the text:
  - Each character is associated a node, and edges between two nodes are created for interactions between those two characters that are explicitly expressed in sentences.
  - Each edge is labeled with the type of interaction:
    - For example *Romeo talked to Juliet* would lead to the creation of an edge between Romeo Montague and Juliet Capulet that is labeled with *talked to*.

- Your solution will be graded based on its precision (how accurate it is) and recall (how many such interactions it correctly extracts).
  - Given that NLP tools are often less than 100\% accurate, elaborate on how you could increase the quality of the extracted social network.

Application: Social Networks from Text

- Pipeline of NLP tasks:
  1. Tokenization and sentence segmentation.
  2. Syntactic parsing.
     - POS tagging to identify pronouns.
  3. Named entity recognition.
  4. Coreference resolution.

Show pipeline processing on the following example:

> Romeo was walking his dog when he met Mercutio's neighbor, Juliet. She asked him if the dog was a golden retriever.

Coreference using SpaCy

- **SpaCy** uses the NeuralCoref model, as implemented by **Huggin Face**:
  - https://spacy.io/universe/project/neuralcoref
  - https://github.com/huggingface/neuralcoref

- Currently not working in version 3 with Python 3.8, need version 2 with Python 3.7:
  - conda create --name spacy-coref
  - conda activate spacy-coref
  - conda install python=3.7
  - pip install spacy==2.1.0
  - pip install neuralcoref
  - pip install https://github.com/explosion/spacy-models/releases//download/en_core_web_lg-2.1.0/en_core_web_lg-2.1.0.tar.gz
  - conda install -c conda-forge jupyterlab # *select Python 3 as kernel in Jupyter.*

Coreference using SpaCy

```
In [1]: # Load your usual SpaCy model (one of SpaCy English models)
        import spacy
        import en_core_web_lg

        nlp = en_core_web_lg.load()

        # Add neural coref to SpaCy's pipe.
        import neuralcoref
        neuralcoref.add_to_pipe(nlp)

        # You can now use NeuralCoref as you usually manipulate a SpaCy document annotations.
        doc = nlp(u'My sister has a dog. She loves him.')

        doc._.has_coref
        doc._.coref_clusters

Out[1]: [My sister: [My sister, She], a dog: [a dog, him]]

In [5]: doc = nlp(u'Romeo was walking his dog when he met Mercutio\'s neighbor, Juliet.')
        doc._.coref_clusters

Out[5]: [Romeo: [Romeo, his, he]]
```

More recent, more accurate approaches

1. CorefQA: Coreference Resolution as Query-based Span Prediction [Wu et al., ACL'20]
   - https://github.com/ShannonAI/CorefQA
2. BERT for Coreference Resolution: Baselines and Analysis [Joshi et al., EMNLP'19]
   - https://github.com/mandarjoshi90/coref
3. Coreference Resolution with Entity Equalization [Kantor & Globerson, ACL'19]
   - https://github.com/kkjawz/coref-ee
4. Higher-order Coreference Resolution with Coarse-to-fine Inference [Lee et al, NAACL'18]
   - https://github.com/kentonl/e2e-coref
5. End-to-end Neural Coreference Resolution [Lee et al., EMNLP'17]
   - https://github.com/kentonl/e2e-coref

Supplementary Readings

- Stanford CS 224N slides on coreference resolution.
- Chapter 21 (Coreference Resolution) in [Jurafsky & Martin]
- Chapter 15 (Reference Resolution) in [Eisenstein]
- The papers on the previous slide.