Project and Paper Presentation

—

Paryul Jain [20171083]
Eesha Dutta [20171104]

Topic and Focus Identification

What is Topic and Focus?

- Topic - what is being talked about
- The part of sentence structure that is presented by the speaker as readily available to the listener's memory
- Focus - what is being asserted about the topic

Practical Application

- Speech Technology - in the design of embodied conversational agents
- Information Retrieval
- Automatic Summarization

Why this problem is not trivial

Topic is not the same as Subject or Agent which can be obtained using syntax and semantics respectively. The actual meaning of the sentence as perceived by a human is required.

1. Topic ≠ Grammatical subject (defined by syntax)

   For example, in the sentence, “As for the little girl, the dog bit her.”

   Subject : “the dog” but Topic : “the little girl”

2. Topic ≠ Agent (defined by semantics)

   For example, in the sentence, “The little girl was bitten by the dog.”

   Agent : “the dog” but Topic: “the little girl”

Why this problem is not trivial

In a discourse, pronouns are used as references to an entity that has already been introduced. We need the antecedent to determine the Topic, and this is not easy.

Anaphora Resolution

- Reference : process by which speakers use expressions to denote an entity
- Referring expression : expression used to perform reference
- Referent : entity being referred to
- Anaphora : reference to an entity previously introduced
- Anaphora Resolution : process of identifying the antecedent to the anaphor
- Anaphora Resolution is classically recognized as a very difficult problem in NLP ([Mitkov 99], [Denber 98])

Constraints (for English)

- Syntactic Constraints:
  - Syntactic relationships between a referring expression and a possible antecedent noun phrase
    - John bought himself a new car. (himself = John)
    - John bought him a new car. (him ≠ John)
- Selectional Restrictions:
  - A verb places restrictions on the anaphor
    - John parked his car in the garage. He had driven it around for hours. (it = car and not garage)
    - I picked up the book and sat in a chair. It broke. (it = chair and not book)

Syntax is not enough

- John hit Bill. He was severely injured. (he = John or Bill?)
- Jane admires Emma and John worships her. (her = Jane or Emma?)

Lappin and Leass Algorithm

Weighing of possible antecedents via recency, syntactic and semantic features

Preferences in Pronoun Interpretation

- Recency
  - Entities introduced recently are more salient than those introduced before
    - John has a bike. Bill has a car. Mike likes to drive it. (it = Mike's car)
- Grammatical Role
  - Entities mentioned in subject position are more salient than those in object position
    - Bill went to the car dealership with John. He bought a Bentley. (he = Bill)
- Repeated Mention
  - Entities that have been focused on in the prior discourse are more salient
    - John needed a car to get to his new job. He decided that he wanted something sporty. Bill went to the car dealership with him. He bought a Bentley. [he = John]

Preferences in Pronoun Interpretation

- Verb Semantics
  - Certain verbs place a semantically-oriented emphasis on one of the arguments
    - John telephoned Bill. He had lost the book in the mall. (he = John)
    - John criticized Bill. He had lost the book in the mall. (he = Bill)
- World Knowledge
    - The city council denied the demonstrators a permit because they feared violence. (they = city council)
    - The city council denied the demonstrators a permit because they advocated violence. (they = demonstrators)

Idea

- Proposes that potential antecedents have degrees of salience
- Try to resolve anaphors by finding highly salient antecedents compatible with pronoun agreement features
- Incorporate
  - Recency
  - Syntax-based preferences
  - Agreement, but no other semantics

Algorithm

- Assign a number of salience factors and salience values to each referring expression

Salience Factor	Salience Value
Sentence recency	100
Subject emphasis	80
Existential emphasis	70
Accusative emphasis	50
Indirect object emphasis	40
Non-adverbial emphasis	50
Head noun emphasis	80

Algorithm

- For each new entity, compute the salience value as the sum of weights assigned by a set of salience factors
- Weights accumulate over time and are cut in half for each sentence processed
- Collect potential referents (up to four sentences back)
- Remove those that don’t agree in number/gender with pronoun
- Select referent with highest salience; it tie, select closest referent in string

Centering Theory

A Framework for Modeling the Local Coherence of Discourse (Grosz, Joshi and Weinstein)

Solve this problem using Centering Theory!!!

- What is discourse coherence?
- Language spoken or written does not contain isolated or unrelated sentences/utterances.
- Rather it contains collocated, structured, coherent groups of sentences.
- Such a coherent structured group of sentences form a discourse.

Example1: John took a train from Paris to Istanbul. Ram likes spinach.
Example2: Jane took a train from Paris to Istanbul. She had to attend a conference.

Which sentence is more coherent? Notion of **inference load** on user.

Methods to approach Discourse Coherence :

1. Relations (Purpose - Attend)
2. Entity (Jane, John)

An Entity Based Coherence Model contrasting with it's counterpart Relation Based Coherence Model.

IDEA : A discourse is coherent if it continues to discuss about the same entity.

A theory of both discourse coherence and discourse salience.

Discourses divide into constituent *discourse segments*.
Local coherence - coherence among the utterances in that segment.
Global coherence - coherence with other segments in the discourse.

(1)

a. John went to his favorite music store to buy a piano.
b. He had frequented the store for many years.
c. He was excited that he could finally buy a piano.
d. He arrived just as the store was closing for the day.

(2)

a. John went to his favorite music store to buy a piano.
b. It was a store John had frequented for many years.
c. He was excited that he could finally buy a piano.
d. It was closing just as John arrived.

Discourse (1) is more coherent than Discourse (2), since in (1) we discuss about the same entity - John. The entity remains constant through the sequence of utterances. But in (2) it flips between John and store, increasing the inference load on user.

Centering theory proves that (1) is more coherent than (2).

Definitions

– **U_n** : an utterance (A sentence or clause with a verb)
– **Center :** An entity that links one utterance to another.
– **Backward-looking center $C_b(U_n)$**: current focus after U_n interpreted.
– **Forward-looking centers $C_f(U_n)$**: ordered list of potential foci referred to in U_n

• $C_b(U_{n+1})$ is highest ranked member of $C_f(U_n)$ realized[1] in U_{n+1}.
• C_f ordered by *subj* > *obj* > *others*.
• $C_p(U_n)$: highest ranked member of $C_f(U_n)$ which is the preferred center of U_{n+1}

The transition between utterances are as follows.

	$C_b(U_{n+1}) = C_b(U_n)$ or $C_b(U_n)$ undef	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

Constraint 1 : For every utterance there is a unique C_b

Rule 1: If any element of $C_f(U_n)$is realized by a pronoun in utterance U_{n+1}, then $C_b(U_{n+1})$ must be realized as a pronoun also.

Rule 2: Transition states are ordered. Continue is preferred to Retain is preferred to Smooth-Shift is preferred to Rough-Shift.

Revisiting the Example

a). John went to his favorite music store to buy a piano. b). He had frequented the store for many years.
c). He was excited that he could finally buy a piano. d). He arrived just as the store was closing for the day.

- $C_f(U_1)$: {John,music store,piano}
- $C_p(U_1)$: John
- $C_b(U_1)$: undefined

- $C_f(U_2)$: {John,store}
- $C_p(U_2)$: John
- $C_b(U_2)$: John

- $C_f(U_3)$: {John,piano}
- $C_p(U_3)$: John
- $C_b(U_3)$: John

- $C_f(U_4)$: {John,store}
- $C_p(U_4)$: John
- $C_b(U_4)$: John

U1 to U2 : CONTINUE
U2 to U3 : CONTINUE
U3 to U4 : CONTINUE

Revisiting the Example

a). John went to his favorite music store to buy a piano. b). It was a store John had frequented for many years. c). He was excited that he could finally buy a piano. d). It was closing just as John arrived.

- $C_f(U_1)$: {John,music store,piano}
- $C_p(U_1)$: John
- $C_b(U_1)$: undefined

- $C_f(U_2)$: {John,store}
- $C_p(U_2)$: John
- $C_b(U_2)$: store

- $C_f(U_3)$: {John,piano}
- $C_p(U_3)$: John
- $C_b(U_3)$: John

- $C_f(U_4)$: {John,store}
- $C_p(U_4)$: John
- $C_b(U_4)$: store

U1 to U2 : ROUGH-SHIFT
U2 to U3 : SMOOTH-SHIFT
U3 to U4 : ROUGH-SHIFT

- In Example 1, there is CONTINUATION of the center, where as in Example 2 there is a ROUGH SHIFT.
- The priority prefers CONTINUATION over ROUGH SHIFT.
- Implies Example 1 is more coherent than Example 2

RESULTS

- **C_b is the closest concept in centering to the traditional notion of Topic** (Sgall 1967; Chafe 1976; Sanford and Garrod 1981;Givon 1983; Vallduvi 1990; Gundel, Hedberg, and Zacharski 1993)
- The rest of the sentence focuses on the topic and hence is the focus.
- Coherence of Discourses can now be compared

We use the above results to solve this issue of anaphora resolution and Topic and Focus identification.

Experiments

Dataset

- News articles - Maharashtra Elections, World Test Championship
- Wikipedia articles - Cricket, Marvel Cinematic Universe
- Research papers - Centering Theory and LDA

Approach

- The data was first cleaned using several preprocessing steps, such as removal of extra spaces, punctuation, numbers and replacing of urls and user mentions with specific token.
- POS tagging, Chunking and Dependency Parsing to obtain Subject, Object and nouns present
- Obtain the referent of pronoun using Feature Specifications
- For every utterance, create a list of C_f and assign C_b
- Topic of the utterance - C_b ; Focus - remaining part
- Discourse Topics are the top few topics present in the Discourse

Results and Analysis

CORPUS	ORIGINAL TOPICS	PREDICTED TOPICS
test1.txt	John, Mike	John, Mike
test2.txt	John, Store	John, Store
test3.txt	MCU	Marvel, franchise, Feige
test4.txt	Research Paper On Centering	Grosz, Discourse, Documents
test5.txt	Origin of Cricket	Cricket, source, claims
test6.txt	Topic and Focus	Topic, Languages, Focus
test7.txt	Latest News	Match, BJP ,Congress
test8.txt	Jack, Jill, Peter, Jane	Jack, Peter
test9.txt	Monkey	Monkey

- Predicted topics very similar to the original ones verifying the correctness of the theory and the implementation. However few errors do creep in.

- The theory's assumption that there is only one topic per sentence by the constraint that there is only one unique C_b per sentence, poses as a limitation for the theory and for the model.

  Example : 'Jack and Jill were going up the hill, while Peter and Jane were coming down. Jack and Jill went up the hill,to fetch a pail of water. Jack fell down and broke his crown, and Jill came tumbling after.'

  The model predicts Jack as the topic, but Jack, Jill, Peter, Jane are all topics of the sentence.

- Another drawback of the implementation is that Centering only picks nouns as the potential centers, which is not the case always again making way for errors.

Alternate Approaches

TOPIC MODELS

Latent Dirichlet Allocation (LDA)

- A Generative Statistical model.
- Discovers abstract topics from a collection of documents.
- Unsupervised Learning
- Documents are Bag of Words
- Converts a **Document*Word** matrix to **Document*Topic** and **Topic*Word** matrix.
- Key Assumption: the way a document was generated was by picking a set of words. So, to find the topics, we reverse engineer the process

Approach

- The data is normalized and cleaned up - Tokenization; Removal of extra spaces, numbers, punctuation marks; Replacing urls and user mentions with special token; Removal of stop words; Lemmatization (reduce to common base form)
- Convert the data into Bag of Words as required by LDA using Gensim
- Train the LDA model and obtain results by varying ‘number of topics’ parameter

Thank You