Topic and Focus Identification

Team Name: Whatever_it_takes
Paryul Jain [20171083]
Eesha Dutta [20171104]

Topic And Focus

In linguistics, the topic of a sentence is what is being talked about, and the focus is what is being said about the topic. Topic (theme, "given" information) can be understood as that part of the sentence structure that is being presented by the speaker as readily available in the listener's memory, whereas the focus (comment, rheme) is what is being asserted about the topic.

Topic, which is defined by pragmatic considerations, is a distinct concept from grammatical **subject**, which is defined by **syntax**. In any given sentence these may be the same, but they need not be. For example, in the sentence "As for the little girl, the dog bit her.", the subject is "the dog" but the topic is "the little girl". Topic and subject are also distinct concepts from **agent** (or actor)—the "doer", which is defined by **semantics**. In English clauses with a verb in the passive voice, for instance, the topic is typically the subject, while the agent may be omitted or may follow the preposition by. For example, in the sentence "The little girl was bitten by the dog.", "the little girl" is the subject and the topic, but "the dog" is the agent.

The sentence- or clause-level "topic" can be defined in a number of different ways. Among the most common are

- the phrase in a clause that the rest of the clause is understood to be about
- a special position in a clause (often at the right or left-edge of the clause) where topics typically appear.

Realization of topic–focus

Different languages mark topics in different ways. Distinct intonation and word-order are the most common means.

In English

The topic/theme comes first in the clause, and is typically marked out by intonation as well.

PAPERS

Topic and Focus

An Automatic Procedure for Topic-Focus Identification

Centering: A Parametric Theory and Its Instantiations

Anaphora Resolution in Discourse

Anaphora : Reference to an entity that has been previously introduced in the discourse.

Anaphora Resolution : Process of identifying the antecedent to the anaphor.

For Example, Ram is tall boy. He is 20 years old. Here “He” is anaphor and its antecedent would be Ram.

Algorithm 1 by Lappin and Leas (‘94)

Weighing via recency and syntactic preferences.

Saliency Factor Weight

1. Sentence recency (in current sentence?) 100
2. Subject emphasis (is it the subject?) 80
3. Existential emphasis (existential prednom?) 70
4. Accusative emphasis (is it the dir obj?) 50
5. Indirect object/oblique comp emphasis 40
6. Non-adverbial emphasis (not in PP,) 50
7. Head noun emphasis (is head noun) 80

Update Rules:

- Weights accumulate over time
- Cut in half after each sentence processed
- Salience values for subsequent referents accumulate for equivalence class of co-referential items (exceptions, e.g. multiple references in same sentence)
- Additional salience weights for grammatical role parallelism (35) and cataphora (-175) calculated when pronoun to be resolved.
- Additional constraints on gender/number agrmt/syntax.

ANAPHORA RESOLUTION :

- Collect potential referents (up to four sentences back):
- Remove those that don’t agree in number/gender with pronoun.
- Remove those that don’t pass intra-sentential syntactic coreference constraints
- The cat washed it.
- Add applicable values for role parallelism (+35) or cataphora (-175) to current salience value for each potential antecedent
- Select referent with highest salience; if tie, select closest referent in string

Algorithm 2 by Grosz ('95) - Centring Theory

Examines interactions between local coherence and the choice of referring expressions
The centers of an utterance are discourse entities serving to link the utterance to other utterances.
Centers are semantic objects, not words, phrases, or syntactic forms but.

- U_n: an utterance
- Backward-looking center $C_b(U_n)$: current focus after Un interpreted
- Forward-looking centers $C_f(U_n)$: ordered list of potential foci referred to in Un
- $C_b(U_{n+1})$ is highest ranked member of $C_f(U_n)$
- C_f may be ordered subj<exist. Prednom<obj<indobj-oblique<dem. advPP (Brennan et al)
- $C_p(U_n)$: preferred (highest ranked) center of $C_f(U_n)$

Transitions from Un to Un+1

	Cb(Un+1)=Cb(Un) or Cb(Un) undef	Cb(Un+1)≠Cb(Un)
Cb(Un+1)= Cp(Un+1)	Continue	Smooth-Shift
Cb(Un+1)≠ Cp(Un+1)	Retain	Rough-Shift

RULES

If any element of $C_f(U_n)$ is pronominalized in U_{n+1}, then $C_b(U_{n+1})$ must also be

Preference: Continue > Retain > Smooth-Shift > Rough-Shift

Algorithm

Generate C_b and C_f assignments for all possible reference assignments

Filter by constraints (syntactic coreference, selectional restrictions,…)

Rank by preference among transition orderings

Example

U_1:George gave Harry a cookie. U_2:He baked the cookie Thursday. U_3: He ate the cookie all up.

- One
  - $C_f(U_1)$: {George,cookie,Harry}
  - $C_p(U_1)$: George
  - $C_b(U_1)$: undefined
- Two
  - $C_f(U_2)$: {George,cookie,Thursday}
  - $C_p(U_2)$: George
  - $C_b(U_2)$: George
  - Continue ($C_p(U_2)$=$C_b(U_2)$; $C_b(U_1)$ undefined

- Three
  - $C_f(U_3)$: {George?,cookie}
  - $C_p(U_3)$: George?
  - $C_b(U_3)$: George?
  - Continue ($C_p(U_3)=C_b(U_3)$; $C_b(U_3)= C_b(U_2)$)
- Or, Three
  - $C_f(U_3)$: {Harry?,cookie}
  - $C_p(U_3)$: Harry?
  - $C_b(U_3)$: Harry?
  - Smooth-Shift ($C_p(U_3)=C_b(U_3)$; $C_b(U_3) \neq C_b(U_2)$)

The winner is…..George!

Definition

Syntactic - of or according to syntax**.**

Salience - the quality of being particularly noticeable or important; prominence.

Semantic - relating to meaning in language or logic.

Coherence - the quality of being logical and consistent.

DOUBTS

1. What data are we using for annotation and also how are we annotating?
2. Did not find any existing tools to implement centering theory on the annotated data. So what is the next step?
3. Either english/hindi?