Introduction to
Natural Language Processing

“Hurt?” “Hurt”

Hiroshi Nakagawa

(Information Technology Center; Mathematical Informatics, Graduate School of Information Science and Technology; Graduate School of Interdisciplinary Information Studies, The University of Tokyo)

nakagawa@dl.itc.u-tokyo.ac.jp

http: //www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/

Pragmatics & Discourse (Ellipsis & Anaphora)

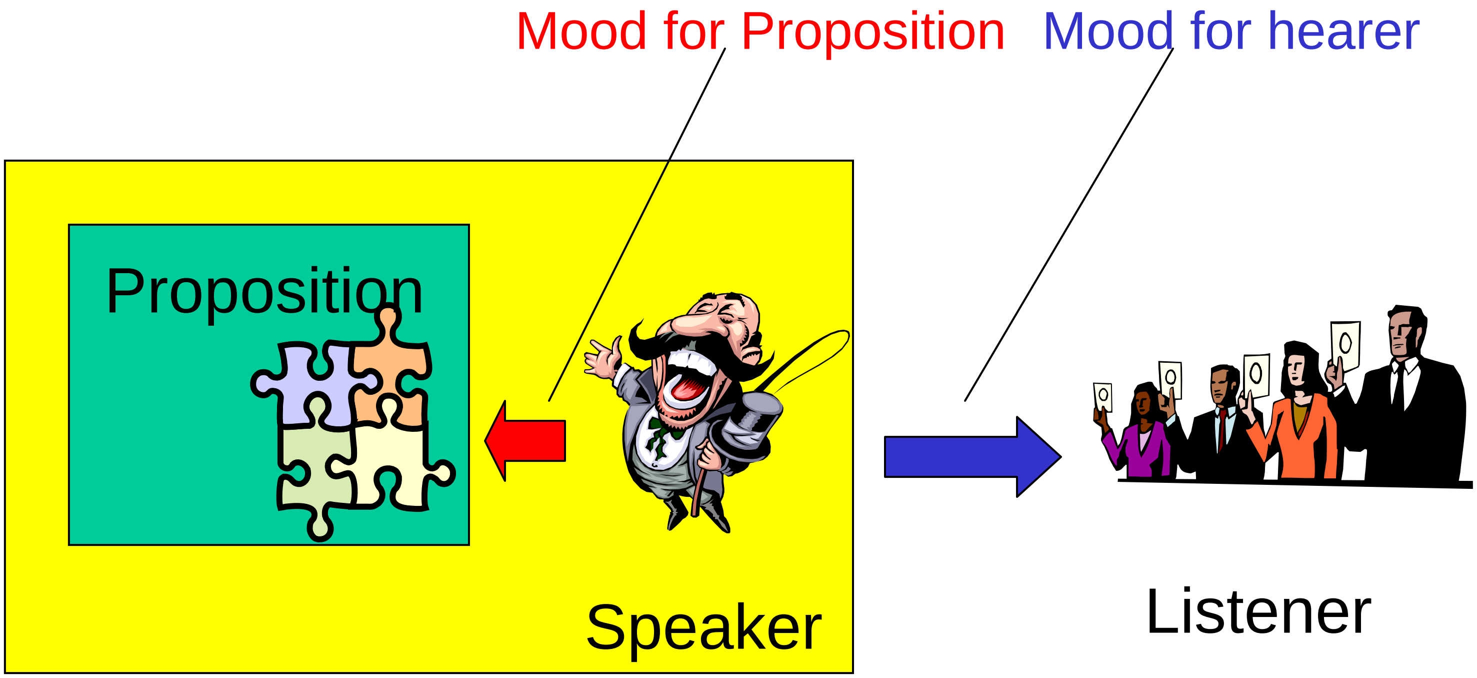
- “Hurt?” = ”(Do you get) hurt?”
- ”Hurt.” = “(I’ve got) hurt.”
- No subject mentioned. Still clear who is hurt.
- This is so called Ellipsis and Anaphora.
- Ellipsis (commonly found in Japanese. Sounds strange if no ellipsis is used.)
    - Other type of Ellipsis: e.g. ”company hereof”, “suspect abovementioned”
- Anaphora: theory to find what a pronoun or an instance of an omission is referring to. In English, the omission of pronoun does not occur. Anaphora resolution is important in finding the reference of pronoun for the understanding of sentence.

Pragmatics & Discourse (Ellipsis & Anaphora)

- Discourse is a term for “a sequence of sentences”. Anaphora is necessary for the understanding of each sentence (and utterance).
- “Frame of reference” is a person whom the speaker is most sympathy for including him/herself.
- ”Hurt” = ”I’ve got hurt”: I (=speaker) is the frame of reference.
- The omission of subject does not diminish the understanding of who. The verb ”hurt” designates the speaker as the frame of reference at default. Such characteristic is seen in some subjective predicates in Japanese (e.g. *ka-yu-I* [itchy]*, ne-mu-i* [sleepy]).
- ”Hurt?” = ”Do you get hurt?”
- ”?” is an expression of asking a question. The speaker does not know him/herself getting hurt. The next candidate for the frame of reference is “you”, the listener.

Theory of Modality (Basics)

- Modality is the theory that describes the mood and intent of the speaker for any instance expressed in a sentence (called ”proposition”).



Mood for Proposition

Mood for hearer

Proposition

Speaker

Listener

Theory of Modality (Basics)

- Modality is the theory that describes the mood and intent of the speaker for any instance expressed in a sentence (called ”proposition”).
- “Hurt.” -> No modality. Proposition (=“hurt”) is directly expressed.
- ”I heard…got hurt./…may be hurt./It is said…got hurt.” -> Quotative evidental mood
- ”…looks like getting hurt” -> Signifies the utterance is based on what the speaker has seen.
- ”…got hurt.” -> Indicates speaker’s intention to express proposition objectively. -> Euphemism
- ”…oh, hurt!” -> Final particle (such as -*oh*[oh] in Japanese) is also a type of modality to listener.

Discourse Anaphora -Centering Theory-

- Examples in Discourse:

1. Taro invited Hanako to go to the movies.
2. *Φ* could not concentrate on anything all day today.

- *Φ* refers to an omitted pronoun called “zero pronoun”. Which does *ϕ* refer to, *Taro* or *Hanako*?
- What algorithm can be applied to rationalize our assumption?
- Hereinafter “Centering Theory (CT)” is discussed, which developed particularly in US during the late 1980s through 1990s.

Centering Theory

- Theory of local discourse coherence (= Degree in which the texts can be considered semantically coherent)
- Unit of discourse = *Utterance U*
- Forward-looking center *Cf (U)*: A set of objects that are referred to in an utterance *U*
- Backward-looking center *Cb (U)*: A central element in *Cf*
- Prominent center *max Cf (U)*: A center in the highest ranked element of *Cf*.
- *Cb* as the center of the current utterance v.s. *Cf* as the center of the next (possible) utterance
- The ordering (ranking) of *Cf* :
    - Topic (subject= -*ha* case) > Frame of reference > -*ga* case > -*ni* case > -*wo* case > other

- Constraints below on utterances *U1,U2,…*:

1. There is precisely one single *Cb (Ui)*.
2. Every element of *Cf (Ui)* must be realized in *Ui* (expressed by a character, zero pronoun, or zero topic (ZTA))
3. *Cb (Ui)* is the highest ranked element of *Cf (Ui-1)*.
4. When some element of *Cf (Ui)* is realized as a pronoun, *Cb (Ui)* is realized as a pronoun in *Ui*.
5. The transition of *Cb* occurs in the following preferred order:

   continue > retain > smooth-shift > rough-shift

	Cb (Ui) = Cb (Ui-1) Or Cb (Ui-1) = Not Specified	Cb (Ui) ≠ Cb (Ui-1)
Cb (Ui) = max Cf (Ui)	continue	smooth-shift
Cb (Ui) ≠ max Cf (Ui)	retain	rough-shift

U1: Taro invited Hanako to go to the movies.
U2: Φ could not concentrate on anything all day today.

- Zero pronoun ϕ in utterance *U* is realized in *U*.

	Cb	Cf	Transition
U1	*Taro*	*Taro (-ga), Hanako (-wo)*, movies *(-ni)*	
U2-a	*Taro*	*Taro (-ga)* as zero pronoun	continue
U2-b	*Hanako*	*Hanako (-ga)* as zero pronoun	smooth-shift

- E.x. continue > retain

1. Taro was invited to a party.
2. ϕ (-ga) liked Hanako very much.
3. I heard ϕ (-ga) invited ϕ (-wo) to go to the movies.

1. Taro was invited to a party.
2. ϕ (-ga) liked Hanako very much.
3. I heard ϕ (-ga) invited ϕ (-wo) to go to the movies.

According to Centering Theory,

1. Cb = *Taro*, Cf = {*Taro*, party},
2. Cb = *Taro*, Cf = {*Taro* (-ga), *Hanako* (-wo)} continue
3. Cb = *Taro*, Cf = {*Taro* (-ga), *Hanako* (-wo)} continue
3. Cb = *Taro*, Cf = {*Hanako* (-ga), *Taro* (-wo)} retain

Compound Sentence

- Anaphora for compound sentence: Resolve compound sentences into simple sentences as a main clause and a subordinate clause. Conjunctive particles denote various phenomena.

  E.g.

  - Group A: φ1 took a train, and (*te*) φ2 went to school. -> φ1=φ2
  - Group B: Because (*no-de*) φ1 went home early, φ2 was saved. -> φ1=?φ2
  - Group C: Although (*ga*) φ1 was expensive, φ2 bought it. -> φ1 ? φ2

Compound Sentence

- Anaphora for compound sentence: Resolve compound sentences into simple sentences as a main clause and a subordinate clause. Conjunctive particles denote various phenomena.

  E.g.

  - (1) Because (*no-de*) ϕ1 had pains, ϕ2 went to bed early. -> ϕ1 = ϕ2
  - ? (2) Because (*no-de*) ϕ1 had pains, ϕ2 went to bed early. -> ϕ1 ≠ ϕ2
  - (3) Because (*no-de*) ϕ1 had pains, …had ϕ2 go to bed early. -> ϕ1 ≠ ϕ2

- Various other factors affects entities.
  - Main clause, predicate in subordinate clause, aspect, tense
  - Currently, natural language processing achieves an app. 80% accuracy by employing anaphora approach.

Rhetorical Structure Theory (RST) (Mann and Thompson)

- Theory of text structure in which the intent of speaker is expressed.
- Rhetorical relations of parts of text in discourse to one another.
  - The parts of text can be clause, sentence, or a sequence of sentences.
- Relations hold between Nucleus and satellite, similar to head-word and subcat in HPSG.
- Meaning yields the organization of texts.

- Text span: part(s) of text for RST
  - Clause as semantic unit
  - Text span consists multiple units.
  - Text span includes a nucleus and satellite.
- Constrains exists between multiple nuclei and satellites.
  - Constrains to nuclei: the reader(s) may not trust nuclei as much as the writer is satisfied.
  - Constrains to satellite: the reader(s) trust satellites.
  - Constrains to the relations holding between nuclei and satellites: the reader(s) trust deeply once he/she understands satellites.
- Effects are the purpose of the writer to the reader(s) by using various RST.
  - Nuclei increase for the reader(s).
  - Nuclei are the primary effects.

Relations between text spans

- Nucleus-satellite Relation
  - Evidence, Justify, Antithesis, Concession, Circumstance, Solutionhood, Elaboration, Background, Enablement, Motivation, (Non) Volitional Cause, (Non) volitional result, Purpose, Condition, Otherwsie, Interpretation, Evaluation, Restatement, Summary, etc. etc.
- Multi-nuclear Relation
  - Sequence, Contrast, Joint, etc.etc.

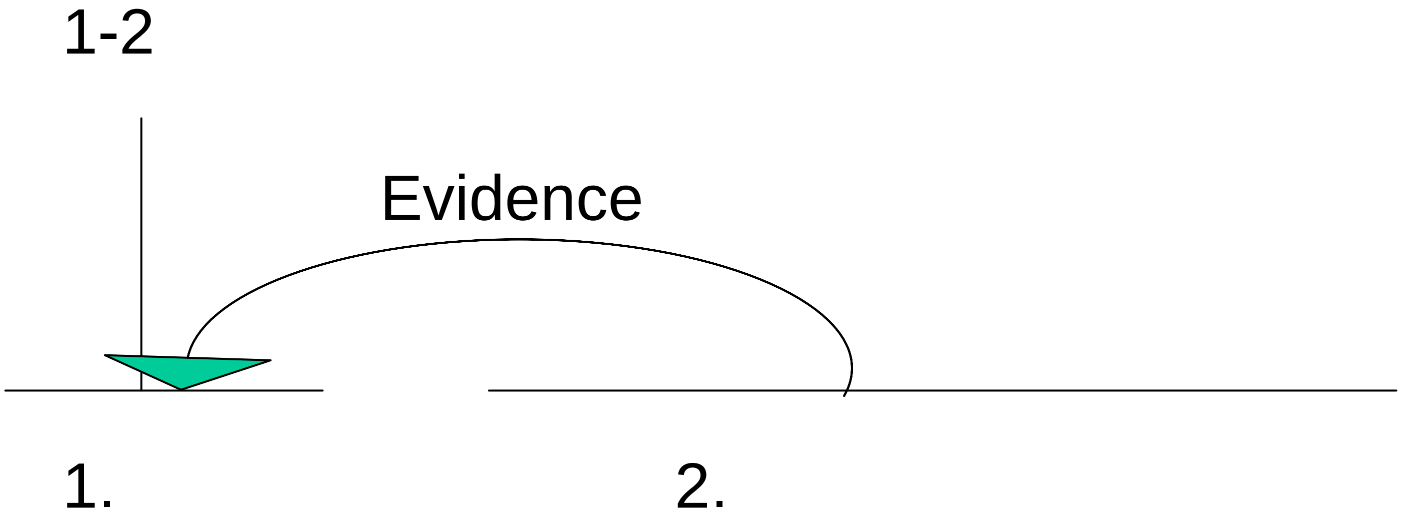
RST Analysis Process

- Step 1. Segment text into units.
  - Text can be segmented into units of an arbitrary size as far as the purpose of RST is met. In many cases, the segmentation of text by clause will lead to an interesting result.
- Step 2. Connect the units to structure text span so that the relations between text spans are clarified.
  - Either “top down” or “bottom up”.
  - The given texts may be defined as multiple relations.

Example

1. The program written for 1980 run properly.

2. The result acquired based on the information in 1980 matched the outcome produced by manual calculation.



Global Structure of Discourse

- Intentional Structure Theory by Groz & Sidner
- The structure of discourse consists:
  - Rhetorical structure = Elements are sequential utterances (unit of discourse).
  - Intentional structure = Purpose of discourse
  - Attentional state = Focus stack
- As the discourse proceeds, the purpose of discourse causes sub purposes to emerge for each unit of discourse.
  - The purpose of discourse (A1) dominates the inferred sub-purpose of discourse (A2).
  - A2 must be satisfied first rather than A1.
- Attentional state is to show the relation between the (sub-)purposes of discourse.

Global Structure of Discourse

- The structure of discourse consists:
  - Rhetorical structure = Elements are sequential utterances (unit of discourse).
  - Intentional structure = Purpose of discourse
  - Attentional state = Focus stack

- The elements of focus are:
  - Elements directly referred to in a discourse unit.
  - Elements referenced to in the discourse unit during the generation and understanding of the discourse unit.
  - (Sub-)purposes of the discourse unit.

Global Structure of Discourse

- Purpose and focus stack structure the discourse.
- Focus stack develops as the discourse proceeds:
    - In Stage 1, DSP1 (the purpose of Discourse Unit 1) is pushed on a stack.
    - In Stage 3, DSP2 (the purpose of Discourse Unit 2) is pushed on the stack.
    - In Stage 7, DSP2 (the purpose of Discourse Unit 2) is popped and cast out from the stack. DSP1 (the purpose of Discourse Unit 1) again comes on top of the stack, meaning DSP1 is the focus of subject.

1	2	3	4	5	6	7	8
DSP1	DSP1	DSP2	DSP2	DSP2	DSP2	DSP1	DSP1
		DSP1	DSP1	DSP1	DSP1		

Key to Grasp the Structure of Discourse

- Change of topic: e.g. a change of *Cb* in Centering Theory
  - Particle *ha* denotes the subject.
- Cue phrase:
  - ”*to-ko-ro-de”* [by the way], ”sa-*te”* [now, well]: Pop on focus stack. A new focus is introduced.
  - "*so-no-ta-me-ni-wa”* [for the purpose], "*ta-to-e-ba”* [for example]: A new focus is introduced. The sub purpose of the discourse is established.
  - "*...si-o-wa-tta-yo”* [(I) have finished up.], “*ko-re-de-OK”* [This is OK.]: The sub purpose is achieved. Pop on focus stack.

Grice’s Theory

- Grice’s “Cooperative Principle” established rules of communication for implicit message by four maxims of conversation.

I. Maxim of quality: Make your statement true.

    I. Do not say what you believe to be false.

    II. Do not say that for which you lack adequate evidence.

II. Maxim of quantity:

    I. Make your contribution to the conversation as informative as necessary.

    II. Do not make your contribution to the conversation more informative than necessary.

III. Maxim of relation: Be relevant.

Grice’s Theory

IV. Maxim of manner:

  I. Avoid obscurity of expression. Avoid ambiguity. Be brief. Be orderly.

- Example:
  - “Five people came yesterday”.
  - Logically speaking, more than five may have come; however, it should be assumed that exactly five people came. Such interpretation is valid if and when the speaker satisfies maxim of quantity and manner.