On the possibility of Machine Translation between UNL dialects: UW aspect

Igor Boguslavsky

Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow / Universidad Politécnica de Madrid

UNL: one language or several dialects?

- UNL varieties
  - – UNL Centre Tokyo
  - – UNDL Geneva
  - – U++ Consortium (France, India, Russia, Spain)
    - Detailed presentation of the U++ position concerning UWs can be found in “UW Guidelines” (to be sent on request)

Initial assumptions

- UWs are labels for complexes of meanings lexicalised in at least some languages.
  - lexicalised = expressed by a single word or non-compositional phrase
- UWs are **language-independent** in the sense that they can denote meanings lexicalised in any language
- UWs are **language-dependent** in the sense that they mostly represent meanings by means of English words.
  - Not simply “English labels” but “English labels + their meaning in English”
  - To some extent, one can modify these meanings by means of constraints

Dialectal differences: U++ vs. UNLC

- Granularity of UWs.
  - UNLC: not (fully) disambiguated UWs are accepted and widespread
    - UW `book` covers all senses of *book*
    - `book(icl>thing)` covers all nominal senses of *book*
  - U++: A UW should refer to one, and only one lexical sense of the word

Constraining UWs

- Semantic constraints should effectively distinguish the meaning we refer to from all other relevant meanings of the headword.
- They should be easily understandable.

Examples

- *Today:* has two senses in English
  - ‘on this day’ (as in: *I am here today but will leave tomorrow*)
  - ‘nowadays’ (as in: *This is no problem today*)
- Therefore UW `today(icl>time)` is insufficient
- Two different UWs needed, e.g.:
  - `today(icl>day>time)`
  - `today(icl>time,equ>nowadays)`

Dialectal differences: U++ vs. UNLC

- Any language for the representation of meaning should effectively express information on the arguments: “who did what to whom”
- UNL (UNLC style) is doing that for verbal concepts:
  - `agt(accuse, minister)` [the minister accused (smb)]
  - `obj(accuse, minister)` [(smb) accused the minister]
- But not for other types of argument-taking concepts
  - *accusation of the minister*:

    `mod(accusation, minister)`
  - *his accusation:*

    `pos(accusation, he)`

Dialectal differences: U++ vs. UNLC

- U++ style:

  `(a) agt(accusation, minister)` [the minister accused smb]

  `(b) obj(accusation, minister)` [smb accused the minister]

- Verbal and nominal predicates should connect their arguments by the same relations
  - `agt(accuse, minister)`
  - `agt(accusation, minister)`
- The distinction between (a) and (b) is important for adequate understanding and question answering.
- E.g. text (a) but not (b) would answer the question *Whom did the minister accuse?*
- UNL `mod(accusation,minister)`does not differentiate between (a) and (b)

Dialectal differences: U++ vs. UNLC

- The information on the arguments a UW can take should be available (their number, the relation they are attached with and the typical semantic class)
- How this information could be represented:
  - constraints within the UW: `write(icl>inform>do,`**`agt>person`**`,`**`obj>uw`**`,`**`rec>person`**`))`
  - a part of the UW description in the UW dictionary

Dialectal differences: U++ vs. UNDL

According to the UNDL style, UWs are represented by the WordNet ID-numbers

- Inconveniences:
  - Unreadable (if the user is not connected to UNDL resources).
  - Does not represent similarities/differences between UWs in the intuitive way. Cf. different but related senses of *girl* that correspond to different synsets:
    - `girl(icl>female) – girl(icl>female_offspring)`
  - No way to restrict the meaning of the English word so that it could be adapted to the Local word meaning
    - Rus. *karij* – `brown(icl>color,aoj>eyes)`
  - No differentiation between meanings expressed by different synset members
  - ID numbers for new concepts should be invented: coordination with Princeton problematic.

Dialectal differences: U++ vs. UNDL

- But maybe there are important advantages that make up for these inconveniences? Possible candidates are:
  - Direct connection to WordNet
  - Disambiguation
- However, U++ style ensures the same:
  - U++ UW dictionary is WN-connected
  - Disambiguation by means of constraints is quite effective – cf. examples in the next slide

Relations used in constraints guarantee easy disambiguation

- `icl, equ, pof, agt, obj,....`
- `ant`
  - `poor(icl>bad):` *poor quality*
  - `poor(ant>rich)`: *poor people*
- A new relation `com` ‘component’ may introduce any relevant meaning component that facilitates disambiguation:
  - `A(com>B)` => B is an important component of the meaning of A

Example

sensational

(a) ‘very good or impressive’: *You look sensational in this dress*

(b) ‘causing intense interest’: *The effect of the discovery was sensational*

UWs

(a) `sensational(icl>good>adj)`

(b) `sensational(icl>adj,com>interest)`

Dialectal differences: attributes

- Traditional view (UNLC and U++): the difference between the UWs and the attributes is related to the **meaning type** (speaker-oriented, modal, pragmatic, etc.). External wrt the concept. Attributes are optional and may be unassigned, if the author does not wish to specify his point of view - the concept remains the same.
- UNDL view (Spec 2010): any meaning may be represented by an attribute. The choice between a UW and an attribute is based on the part of speech of the underlying NL word
  - Only N, V, Adj, Adv can generate UWs.
  - Any meaning expressed by a Pr/Conj in at least one NL loses the right to be expressed by a UW and should generate an attribute or a relation

Inconveniences of the UNDL view

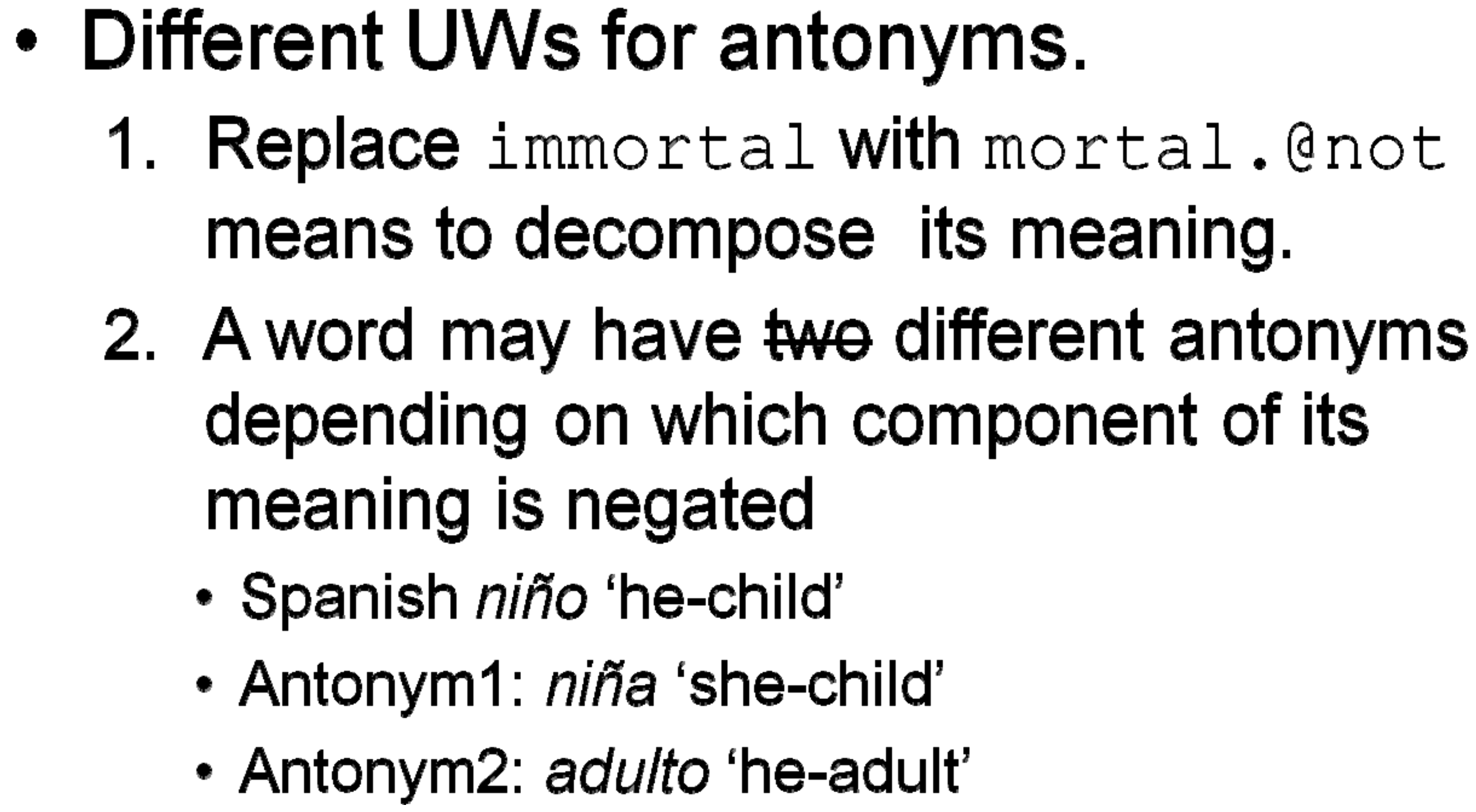
- A concept can be realized both as an open class word and a closed class word in the same language (*to <u>cause</u> – (die) <u>of</u> (hunger), <u>from</u> (starvation)*)
- UNDL: any meaning can be made an attribute:
  - *to hunger* = `hunger@full_of.@make`

This contradicts the following important postulate about UWs which concerns their granularity.

UW dictionary is a collection of lexicalized concepts of all languages

- A UW should have a one-word equivalent in at least one language. The decision wrt UWs is taken depending on what kind of words exist in NLs.
- **NO lexical meaning decomposition.** UWs disambiguate NL words but do not define their meaning.
- If we begin decomposing the lexical meaning of some words (*to hunger* = ‘make somebody full of hunger’), we should do it consistently and decompose them all. This will be an entirely different project.
- This answers Question 3.

Question 4: antonyms



- **Different UWs for antonyms.**
  1. **Replace** `immortal` **with** `mortal.@not` **means to decompose its meaning.**
  2. A word may have ~~two~~ different antonyms depending on which component of its meaning is negated
     - Spanish *niño* 'he-child'
     - Antonym1: *niña* 'she-child'
     - Antonym2: *adulto* 'he-adult'

Question 5: multiword expressions

- The important distinction is:
  - not between “a NL word” vs. “a NL phrase”
  - but between “a compositional phrase” and “a word or a non-compostional phrase that denotes a single concept”
- If the phrase is compositional : no UW
- If there is a word or a non-compositional phrase: a UW. Options:
  - Simple UW (if exists in English)
  - Multiword headword (`cable_railway(icl>transport)`)
  - Hypernode (`(mod(railway,cable)(icl>transport)`)

Question 1: *most*

- If we wish to make inferences based on UNL graphs, we should treat *most* as a 3-place predicate: `most(X,Y,Z)`= ‘X has property Y in a greater degree than any other element of set Z does’

(1) *The most interesting* (Y) *paper* (X) *on the program* (Z)

- Arguments X, Y and Z are needed for understanding the *most* situation*.* Since attributes cannot take arguments, *most* should be a UW.
- Prepositions *on, of, among,* that introduce argument Z, should be omitted from the graph.
- Superlatives should be represented by means of *most: the greatest* – `‘the most great’`

Question 1: *generally regarded as*

Sentences (1) - (4) contain the same verbal concept:

- *He is generally regarded as a great writer*
- *He is regarded by all as a great writer*
- *He is regarded by us as a great writer*
- *We regard him as a great writer*
- Hence, two UWs needed:
  - *regarded as* → `regard`
  - *generally* → `all`

Question 2: *Charles Dickens*

- Two interpretations:
  - ‘a person whose name is Charles Dickens’
  - ‘a famous English novelist whose name is Charles Dickens’
- For both interpretations, it is convenient to have special dictionaries, but with different amount of information
- 1st interpretation: A dictionary of proper names
- 2nd interpretation refers to the background knowledge: A dictionary of individuals.
  - Requires much more elaborated structure

To sum up:

- Bridging the gap between UNL dialects is useful and, hopefully, possible.
- Major differences concern:
  - UWs:
    - Ambiguity allowed/not
    - Decomposition allowed/not
    - Information on arguments given/not
    - Noun-Verb argument structures parallel/not
  - Nature of attributes
    - Speaker-oriented/any meaning

What can be done?

- Organise technical consultations aiming at overcoming the differences between the dialects or finding a way to establish a correspondence between them.
- Set up a common database which would represent all existing UW dictionaries and establish links between them.
  - Computational support for such a database already exists (PIVAX system, Grenoble).