CS712: Interlingua Based MT

Pushpak Bhattacharyya
CSE Department
IIT Bombay
28/1/13
Empiricism vs. Rationalism

- Ken Church, “A Pendulum Swung too Far”, LILT, 2011
  - Availability of huge amount of data: what to do with it?
  - 1950s: Empiricism (Shannon, Skinner, Firth, Harris)
  - 1970s: Rationalism (Chomsky, Minsky)
  - 1990s: Empiricism (IBM Speech Group, AT & T)
  - 2010s: Return of Rationalism?
Introduction

• Machine Translation (MT) is a technique to translate texts from one natural language to another natural language using a machine.

• Translated text should have two desired properties:
  – Adequacy: Meaning should be conveyed correctly
  – Fluency: Text should be fluent in the target language

• Translation between distant languages is a difficult task
  – Handling Language Divergence is a major challenge
Kinds of MT Systems
(point of entry from source to the target text)

Deep understanding level
Interlingual level
Logico-semantic level
Mixing levels
Syntactico-functional level
Syntagmatic level
Morpho-syntactic level
Graphemic level

Direct translation
Semi-direct translation
Multilevel transfer
Syntactic transfer (surface)
Syntactic transfer (deep)
Semantic transfer
Conceptual transfer

Ontological interlingua
Semantico-linguistic interlingua
SPA-structures (semantic & predicate-argument)
Multilevel description
F-structures (functional)
C-structures (constituent)
Tagged text
Text
Why is MT difficult: Language Divergence

• One of the main complexities of MT: 
  *Language Divergence*

• Languages have different ways of expressing meaning
  – Lexico-Semantic Divergence
  – Structural Divergence

English-IL Language Divergence with illustrations from Hindi
*(Dave, Parikh, Bhattacharyya, Journal of MT, 2002)*
Language Divergence Theory:

*Lexico-Semantic Divergences*

- Conflational divergence
  - F: vomir; E: to be sick
  - E: *stab*; H: *churaa se maaranaa* (*knife-with hit*)
  - S: *Utrymningsplan*; E: *escape plan*
- Structural divergence
  - E: *SVO*; H: *SOV*
- Categorial divergence
  - Change is in POS category (many examples discussed)
Language Divergence Theory: Lexico-Semantic Divergences (cntd)

- Head swapping divergence
  - E: Prime Minister of India; H: bhaarat ke pradhaan mantrii (India-of Prime Minister)
- Lexical divergence
  - E: advise; H: paraamarsh denaa (advice give): Noun Incorporation- very common Indian Language Phenomenon
Language Divergence Theory: 
*Syntactic Divergences*

- **Constituent Order divergence**
  - **E:** Singh, the PM of India, will address the nation today; **H:** bhaarat ke pradhaan mantrii, singh, … *(India-of PM, Singh…)*

- **Adjunction Divergence**
  - **E:** She will visit here in the summer; **H:** vah yahaa garmii meM aayegii (she here summer-in will come)

- **Preposition-Stranding divergence**
  - **E:** Who do you want to go with?; **H:** kisake saath aap jaanaa chaahate ho? (who with…)
Language Divergence Theory: 
**Syntactic Divergences**

- **Null Subject Divergence**
  - E: I will go; H: jaauMgaa (subject dropped)

- **Pleonastic Divergence**
  - E: It is raining; H: baarish ho rahii haai (rain happening is: no translation of it)
Language Divergence exists even for close cousins
(Marathi-Hindi-English: case marking and postpositions do not transfer easily)

• संनिहित भूत (immediate past)
  – कधी आलास? हा येतो इतकाच! (M)
  – कब आये? बस अभी आया! (H)
  – When did you come? Just now (I came) (H)
  – kakhana ele? ei elaam/eshchhi (B)

• नि:संशय भविष्य (certainty in future)
  – आता तो मार खातो खास! (M)
  – अब वह मार खायगा ही! (H)
  – He is in for a thrashing. (E)
  – ekhan o maar khaabei/khaachhei

• आश्वासन (assurance)
  – मी तुम्हाला उद्या भेटतो. (M)
  – मैं आप से कल मिलता हूँ! (H)
  – I will see you tomorrow. (E)
  – aami kaal tomaar saathe dekhaa korbo/korchhi (B)
Interlingua Based MT
KBMT


• Forerunner of many interlingua based MT efforts
IL based MT: schematic
Specification of KBMT

- Source languages: English and Japanese
- Target languages: English and Japanese
- Translation paradigm: Interlingua;
- Computational architecture: A distributed, coarsely parallel system
- Subworld (domain) of translation: personal computer installation and maintenance manuals.
Knowledge base of KBMT (1/2)

- An ontology (domain model) of about 1,500 concepts
- Analysis lexicons: about 800 lexical units of Japanese and about 900 units of English
- Generation lexicons: about 800 lexical units of Japanese and about 900 units of English
Knowledge base of KBMT (2/2)

- Analysis grammars for English and Japanese;
- Generation grammars for English and Japanese;
- Specialized syntax-semantics structural mapping rules;
Underlying formalisms

- The knowledge representation system FrameKit
- A language for representing domain models (a semantic extension of FRAMEKIT)
- Specialized grammar formalisms, based on Lexical-Functional Grammar
- A specially constructed language for representing text meanings (the interlingua)
- The languages of analysis and generation lexicon entries, and of the structural mapping rules
Procedural components

• A syntactic parser with a semantic constraint interpreter
• A semantic mapper for treating additional types of semantic constraints
• An interactive augmentor for treating residual ambiguities;
Architecture of KBMT
Digression: language typology
Language and dialects

• There are 6909 living languages (2009 census)
• Dialects far outnumber the languages
• Language varieties are called dialects
  – if they have no standard or codified form,
  – if the speakers of the given language do not have a state of their own,
  – if they are rarely or never used in writing (outside reported speech),
  – if they lack prestige with respect to some other, often standardised, variety.
“Linguistic” interlingua

• Three notable attempts at creating interlingua
  – “Interlingua” by IALA (International Auxiliary Language Association)
  – “Esperanto”
  – “Ido”
The Lord’s Prayer in Esperanto

Interlingua version

1. Patro Nia, kiu estas en la ĉielo, via nomo estu sanktigita.
2. Venu via regno, plenumiĝu via volo, kiel en la ĉielo, tiel ankaŭ sur la tero.
3. Nian panon ĉiutagan donu al ni hodiaŭ.
5. Kaj ne konduku nin en tenton, sed liberigu nin de la malbono.
Amen.

Latin version

1. Pater noster, qui es in caelis, sanctificetur nomen tuum.
2. Adveniat regnum tuum.
3. Fiat voluntas tua, sicut in caelo, et in terra.
5. E et dimittte nobis debita nostra, sicut et nos dimittimus debitoribus nostris.
6. Et ne nos inducas in tentationem, sed libera nos a malo.
Amen.

English version (traditional)

1. Our Father, who art in heaven, hallowed be thy name;
2. thy kingdom come, thy will be done.
3. on earth, as it is in heaven.
4. Give us this day our daily bread;
5. and forgive us our debts as we have forgiven our debtors.
6. And lead us not into temptation, but deliver us from evil.
Amen.
“Interlingua” and “Esperanto”

- *Control languages* for “Interlingua”: French, Italian, Spanish, Portuguese, English, German and Russian
- Natural words in “interlingua”
- “manufactured” words in Esperanto, using heavy agglutination
- Esperanto word for "hospital“: mal·san·ul·ej·o: *mal* (opposite), *san* (health), *ul* (person), *ej* (place), *o* (noun)
Twelve Extant Phyla of Human Language

- Australian
- Bantu, etc. (Congo-Kordofanian)
- Caucasian
- Dravidian

Additional Phyla:
- Eurasian
- East Asian
- Georgain (Kartvelian)
- Hmong, etc. (Austro-Asiatic)
- Indo-Pacific
- Jivaroan, etc. (Amerind)
- Khoisan
- Nilotic-Saharan
Language Universals vs. Language Typology

• “Universals” is concerned with what human languages have in common, while the study of typology deals with ways in which languages differ from each other.
Typology: basic word order

- SOV (Japanese, Tamil, Turkish etc.)
- SVO (Fula, Chinese, English etc.)
- VSO (Arabic, Tongan, Welsh etc.)

“Subjects tend strongly to precede objects.”
Typology: Morphotactics of singular and plural

- No expression: Japanese *hito* 'person', pl. *hito*
- Function word: Tagalog *bato* 'stone', pl. *mga bato*
- Affixation: Turkish *ev* 'house', pl. *ev-ler*; Swahili *m-toto* 'child', pl. *wa-toto*
- Sound change: English *man*, pl. *men*; Arabic *rajulun* 'man', pl. *rijalun*
- Reduplication: Malay *anak* 'child', pl. *anak-anak*
Typology: Implication of word order

Due to its SVO nature, English has:
- preposition+noun (in the house)
- genitive+noun (Tom's house) or noun+genitive (the house of Tom)
- auxiliary+verb (will come)
- noun+relative clause (the cat that ate the rat)
- adjective+standard of comparison (better than Tom)
Typology: motion verbs (1/3)

• Motion is expressed differently in different languages, and the differences turn out to be highly significant.

• There are two large types:
  – verb-framed and
  – satellite-framed languages.
Typology: motion verbs (2/3)

- In satellite-framed languages like English, the motion verb typically also expresses **manner** or **cause**:
  - *The bottle floated out of the cave.* (Manner)
  - *The napkin blew off the table.* (Cause)
  - *The water rippled down the stairs.* (Manner)
Typology: motion verbs (3/3)

• Verb-framed languages express manner and cause, not in the verb, but in a more peripheral element:

• Spanish
  – La botella salió de la cueva flotando.
  – The bottle exited the cave floatingly

• Japanese
  – Bin-ga dookutsu-kara nagara-te de -ta
  – bottle-NOM cave-from float-GER exit-PAST
A look at annotation
Definition etc. (1/2)

(Eduard Hovy, ACL 2010, tutorial on annotation)

• Annotation (‘tagging’) is the process of adding new information into raw data by humans annotators.

• Usually, the information is added by many small individual decisions, in many places throughout the data.

• The addition process usually requires some sort of mental decision that depends both on the raw data and on some theory or knowledge that the annotator has internalized earlier.
Definition etc. (1/2)

• Typical annotation steps:
  – Decide which fragment of the data to annotate
  – Add to that fragment a specific bit of information
  – chosen from a fixed set of options
(According to a new research, those people who have a busy social life, have larger space in a part of their brain).

(Neuroscientists have found that people who have a busy social life have larger space in a part of their brain.)
Ambiguity of लोगों (People)

- लोग, जन, लोक, जनमानस, पब्लिक - एक से अधिक व्यक्ति "लोगों के हित में काम करना चाहिए"
  - (English synset)
    multitude, masses, mass, hoi_polloi, people, the_great_unwashed - the common people generally "separate the warriors from the mass" "power to the people"

- दुनिया, दुनियाँ, संसार, विश्व, जगत, जहाँ, जहान, ज़माना, जमाना, लोक, दुनियावाले, दुनियाँवाले, लोग - संसार में रहने वाले लोग "महात्मा गाँधी का सम्मान पूरी दुनिया करती है / में इस दुनिया की परवाह नहीं करता / आज की दुनिया पैसे के पीछे भाग रही है"
  - (English synset) populace, public, world - people in general considered as a whole "he is a hero in the eyes of the public"
Sense Marked corpora in Marathi

Snapshot of a Marathi sense tagged paragraph
A good annotator is rare to find!

- An annotator has to understand BOTH language phenomena and the data
- An annotation designer has to understand BOTH linguistics and statistics!

Linguistics and Language phenomena

Annotator

Data and statistical phenomena
The completeness chain

- MT is NLP-complete
- NLP is AI-complete
- AI id CS-complete
Penn tagset (1/2)

<table>
<thead>
<tr>
<th>CC</th>
<th>Coord Conjunct</th>
<th>and, but, or</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>Cardinal number</td>
<td>one, two</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
<td>the, some</td>
</tr>
<tr>
<td>EX</td>
<td>Existential there</td>
<td>there</td>
</tr>
<tr>
<td>FW</td>
<td>Foreign Word</td>
<td>mon dieu</td>
</tr>
<tr>
<td>IN</td>
<td>Preposition</td>
<td>of, in, by</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
<td>big</td>
</tr>
<tr>
<td>JJR</td>
<td>Adj., comparative</td>
<td>bigger</td>
</tr>
<tr>
<td>JJS</td>
<td>Adj., superlative</td>
<td>biggest</td>
</tr>
<tr>
<td>LS</td>
<td>List item marker</td>
<td>1, One</td>
</tr>
<tr>
<td>MD</td>
<td>Modal</td>
<td>can, should</td>
</tr>
<tr>
<td>NN</td>
<td>Noun, sing. or mass</td>
<td>dog</td>
</tr>
<tr>
<td>NNS</td>
<td>Noun, plural</td>
<td>dogs</td>
</tr>
<tr>
<td>NNP</td>
<td>Proper noun, sing.</td>
<td>Edinburgh</td>
</tr>
<tr>
<td>NNPS</td>
<td>Proper noun, plural</td>
<td>Orkneys</td>
</tr>
<tr>
<td>PDT</td>
<td>Predeterminer</td>
<td>all, both 's</td>
</tr>
<tr>
<td>POS</td>
<td>Possessive ending</td>
<td>I, you, she</td>
</tr>
<tr>
<td>PP</td>
<td>Personal pronoun</td>
<td>my, one’s</td>
</tr>
<tr>
<td>PP$</td>
<td>Possessive pronoun</td>
<td>quickly</td>
</tr>
<tr>
<td>RB</td>
<td>Adverb</td>
<td>faster</td>
</tr>
<tr>
<td>RBR</td>
<td>Adverb, comparative</td>
<td>fastest</td>
</tr>
</tbody>
</table>
Penn tagset (2/2)

<table>
<thead>
<tr>
<th>RP</th>
<th>Particle</th>
<th>up, off</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYM</td>
<td>Symbol</td>
<td>+, %, &amp;</td>
</tr>
<tr>
<td>TO</td>
<td>“to”</td>
<td>to</td>
</tr>
<tr>
<td>UH</td>
<td>Interjection</td>
<td>oh, oops</td>
</tr>
<tr>
<td>VB</td>
<td>verb, base form</td>
<td>eat</td>
</tr>
<tr>
<td>VBD</td>
<td>verb, past tense</td>
<td>ate</td>
</tr>
<tr>
<td>VBG</td>
<td>verb, gerund</td>
<td>eating</td>
</tr>
<tr>
<td>VBN</td>
<td>verb, past part</td>
<td>eaten</td>
</tr>
<tr>
<td>VBP</td>
<td>Verb, non-3sg, pres</td>
<td>eat</td>
</tr>
<tr>
<td>VBZ</td>
<td>Verb, 3sg, pres</td>
<td>eats</td>
</tr>
<tr>
<td>WDT</td>
<td>Wh-determiner</td>
<td>which, that</td>
</tr>
<tr>
<td>WP</td>
<td>Wh-pronoun</td>
<td>what, who</td>
</tr>
<tr>
<td>WP$</td>
<td>Possessive-Wh</td>
<td>whose</td>
</tr>
<tr>
<td>WRB</td>
<td>Wh-adverb</td>
<td>how, where</td>
</tr>
<tr>
<td>$</td>
<td>Dollar sign</td>
<td>$</td>
</tr>
<tr>
<td>#</td>
<td>Pound sign</td>
<td>#</td>
</tr>
<tr>
<td>&quot;</td>
<td>Left quote</td>
<td>', '&quot;</td>
</tr>
<tr>
<td>&quot;</td>
<td>Right quote</td>
<td>', '&quot;</td>
</tr>
<tr>
<td>(</td>
<td>Left paren</td>
<td>(</td>
</tr>
<tr>
<td>)</td>
<td>Right paren</td>
<td>)</td>
</tr>
<tr>
<td>,</td>
<td>Comma</td>
<td>,</td>
</tr>
<tr>
<td>:</td>
<td>Sent-final punct</td>
<td>! ?</td>
</tr>
<tr>
<td>:</td>
<td>Mid-sent punct.</td>
<td>; — ...</td>
</tr>
</tbody>
</table>
## Indian Language Tagset: Noun

<table>
<thead>
<tr>
<th>Sl. No</th>
<th>Category</th>
<th>Label</th>
<th>Annotation Convention</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Noun</td>
<td>N</td>
<td>N</td>
<td>ladakaa, raajaa, kitaaba</td>
</tr>
<tr>
<td>1.1</td>
<td>Common</td>
<td>NN</td>
<td>N__NN</td>
<td>kitaaba, kalam, cashmaa</td>
</tr>
<tr>
<td>1.2</td>
<td>Proper</td>
<td>NNP</td>
<td>N__NNP</td>
<td>Mohan, ravi, rashmi</td>
</tr>
<tr>
<td>1.4</td>
<td>Nloc</td>
<td>NST</td>
<td>N__NST</td>
<td>Uupara, niice, aage,</td>
</tr>
</tbody>
</table>
Indian Language Tagset: Verb, Adjective, Adverb

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Verb</td>
<td>V</td>
<td>V</td>
<td>giraa, gayaa, sonaa, haMstaa, hai, raha</td>
</tr>
<tr>
<td>4.1</td>
<td>Main</td>
<td>VM</td>
<td>V__VM</td>
<td>giraa, gayaa, sonaa, haMstaa,</td>
</tr>
<tr>
<td>4.2</td>
<td>Auxiliary</td>
<td>VAUX</td>
<td>V__VAUX</td>
<td>hai, raha, huaa,</td>
</tr>
<tr>
<td>5</td>
<td>Adjective</td>
<td>JJ</td>
<td>JJ</td>
<td>sundara, acchaa, baRaa</td>
</tr>
<tr>
<td>6</td>
<td>Adverb</td>
<td>RB</td>
<td>RB</td>
<td>jaldii, teza</td>
</tr>
</tbody>
</table>
## Indian Language Tagset: Particle

<table>
<thead>
<tr>
<th></th>
<th>Particles</th>
<th>RP</th>
<th>RP</th>
<th>to, bhii, hii</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Default</td>
<td>RPD</td>
<td>RP_RPD</td>
<td>to, bhii, hii</td>
</tr>
<tr>
<td>9.1</td>
<td>Interjection</td>
<td>INJ</td>
<td>RP_INJ</td>
<td>are, he, o</td>
</tr>
<tr>
<td>9.4</td>
<td>Intensifier</td>
<td>INTF</td>
<td>RP_INTF</td>
<td>bahuta, behada</td>
</tr>
<tr>
<td>9.5</td>
<td>Negation</td>
<td>NEG</td>
<td>RP_NEG</td>
<td>nahiin, mata, binaa</td>
</tr>
</tbody>
</table>
Scale of effort involved in annotation: Penn Treebank (1/2)

- Wall Street Journal subcorpus of the PTB 2 release
- 8 people working half time each about a year
- The same group had previously worked on annotation guidelines, and followed yet another year of an initial annotation of the WSJ corpus that was redone for the PTB 2 version
The same group also
- annotated most of the original Brown corpus with PTB 2 style trees
- 1 million words of the Switchboard corpus, for total of well over 3 million words
- POS tagged 3 million words of a much larger WSJ corpus

All in all, the effort was over about 5 years, and had about 4-5 full time employees.

Total effort: **8 million words, 20-25 man years**
Annotation effort: Ontonotes

• Annotate 300K words per year
  – comprising various genres of text (news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows)
  – in three languages (English, Chinese, and Arabic)
  – with structural information (syntax and predicate argument structure) and
  – shallow semantics (word sense linked to an ontology and coreference).

• 1 person (Dr. Ann Taylor, senior lecturer in Yor Univ), had managed the treebank project for a year, in the last stages of the project

• Ontonotes annotation guidelines were frozen a priori
Annotation effort: Prague Discourse Treebank (1/2)

- Phase 1: autumn 2007 through September 2009: collecting information, getting acquainted with the annotation scheme of the Pennsylvania Discourse Treebank. Development of the annotation scenario (a team of three or four people)
- Phase 2: from September 2009 till January 2010: 5 (part time) annotators, testing our scheme
Annotation effort: Prague Discourse Treebank (2/2)

• Phase 3: from January 2010 till May 2012: 5 annotators part time annotating the PDT (about 50000 sentences in more than 3000 documents)

• Phase 4: from May 2012 till October 2012: testing for consistency, corrections etc.
Back to Interlingua
MT: EnConversion + Deconversion
Challenges of interlingua generation

- Mother of NLP problems - Extract meaning from a sentence!
- Almost all NLP problems are sub-problems
  - Named Entity Recognition
  - POS tagging
  - Chunking
  - Parsing
  - Word Sense Disambiguation
  - Multiword identification
  - and the list goes on...
UNL: a United Nations project

- Started in 1996
- 10 year program
- 15 research groups across continents
- First goal: generators
- Next goal: analysers (needs solving various ambiguity problems)
- Current active language groups
  - UNL_French (GETA-CLIPS, IMAG)
  - UNL_Hindi (IIT Bombay with additional work on UNL_English)
  - UNL_Italian (Univ. of Pisa)
  - UNL_Portuguese (Univ of Sao Paolo, Brazil)
  - UNL_Russian (Institute of Linguistics, Moscow)
  - UNL_Spanish (UPM, Madrid)
World-wide
Universal Networking Language (UNL) Project

- Language independent meaning representation.
UNL represents knowledge:

*John eats rice with a spoon*

**Diagram:**

- **eat(icol>do)**
  - **agt**: John(icol>person)
  - **ins**: rice(icol>food)
  - **obj**: spoon(icol>artifact)
  - **@entry. @present**

**Labels:**

- Universal words
- Semantic relations
- Attributes

**Repository:**

- 42 Semantic Relations
- 84 Attribute Labels
Deepa claimed that she had composed a poem.

[UNL]

agt(claim.@entry.@past, Deepa)
obj(claim.@entry.@past, :01)
agt:01(compose.@past.@entry.@complete, she)
obj:01(compose.@past.@entry.@complete, poem.@indef)

[\UNL]
The Lexicon

He **forwarded** the **mail** to the **minister**.

Content words:

[forward] {} “forward(icl>send)” (V,VOA) <E,0,0>;

[mail] {} “mail(icl>message)” (N,PHSCL,INANI) <E,0,0>;

[minister] {} “minister(icl>person)” (N,ANIMT,PHSCL,PRSN) <E,0,0>;

Headword → Universal Word → Attributes
The Lexicon (cntd)

He forwarded the mail to the minister.

function words:

[he] {} “he” (PRON, SUB, SING, 3RD) <E, 0, 0>;

[the] {} “the” (ART, THE) <E, 0, 0>;

[to] {} “to” (PRE, #TO) <E, 0, 0>;

Headword Universal Word Attributes
How to obtain UNL expressions

• UNL nerve center is the verb
• English sentences:
  – A <verb> B or
  – A <verb>

Recurse on A and B
Enconversion

Long sequence of masters theses: Anupama, Krishna, Sandip, Abhishek, Gourab, Subhajit, Janardhan in collaboration with Rajat, Jaya, Gajanan, Rajita

Many publications
System Architecture

NER

Stanford Dependency Parser

Clause Marker

WSD

Simplifier


Merger

Simple Sentence Enconverter

Stanford Dependency Parser

XLE Parser

Feature Generation

Attribute Generation

Relation Generation
Complexity of handling long sentences

- **Problem**
  - As the length (number of words) increases in the sentence the XLE parser fails to capture the dependency relations precisely
  
  - The UNL enconversion system is tightly coupled with XLE parser

- Resulting in fall of accuracy.
Solution

• Break long sentences to smaller ones
  – Sentence can only be broken at the clausal levels: Simplification
  – Once broken they needs to be joined: Merging

• Reduce the reliance of the Encoverter on XLE parser
  – The rules for relation and attribute generation are formed on Stanford Dependency relations
Text Simplifier

• It simplifies a complex or compound sentence.

• Example:
  – Sentence: John went to play and Jane went to school.
  – Simplified: John went to play. Jane went to school.

• The new simplifier is developed on the clause-markings and the inter-clause relations provided by the Stanford dependency parser
UNL Merging

• Merger takes multiple UNLs and merges them to form one UNL.

• Example-
  – Input:
    • \texttt{agt(go@entry, John), pur(go@entry, play)}
    • \texttt{agt(go@entry, Mary), gol(go@entry, school)}
  – Output:
    • \texttt{agt(go@entry, John), pur(go@entry, play)}
    • \texttt{and (go, :01)}
    • \texttt{agt:01(go@entry, Mary), gol:01(go@entry, school)}

• There are several cases based on how the sentence was simplified. Merger uses rules for each of the cases to merge them.
NLP tools and resources for UNL generation

Tools

- Stanford Named Entity Recognizer (NER)
- Stanford Dependency Parser
- Word Sense Disambiguation (WSD) system
- Xerox Linguistic Environment (XLE) Parser

Resource

- Wordnet
- Universal Word Dictionary (UW++)
System: Processing Units

**Syntactic Processing**
- NER
- Stanford Dependency Parser
- XLE Parser

**Semantic Processing**
- Stems of words
- WSD
- Noun originating from verbs
- Feature Generation
  - Noun features
  - Verb features
SYNTACTIC PROCESSING
Syntactic Processing: NER

- NER tags the named entities in the sentence with their types.

- Types covered
  - Person
  - Place
  - Organization

- Input: *Will Smith was eating an apple with a fork on the bank of the river.*

- Output: `<PERSON>Will Smith</PERSON> was eating an apple with a fork on the bank of the river.`
Syntactic Processing: Stanford Dependency Parser (1/2)

- Stanford Dependency Parser parses the sentence and produces
  - POS tags of words in the sentence
  - Dependency parse of the sentence
**Syntactic Processing: Stanford Dependency Parser (2/2)**

<table>
<thead>
<tr>
<th>Input</th>
<th>Will-Smith was eating an apple with a fork on the bank of the river.</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS tags</td>
<td>Will-Smith/NNP was/VBD eating/VBG an/DT apple/NN with/IN a/DT fork/NN on/IN the/DT bank/NN of/IN the/DT river/NN ./</td>
</tr>
</tbody>
</table>

**Dependency Relations**

- nsubj(eating-3, Will-Smith-1)
- aux(eating-3, was-2)
- det(apple-5, an-4)
- dobj(eating-3, apple-5)
- prep(eating-3, with-6)
- det(fork-8, a-7)
- pobj(with-6, fork-8)
- prep(eating-3, on-9)
- det(bank-11, the-10)
- pobj(on-9, bank-11)
- prep(bank-11, of-12)
- det(river-14, the-13)
- pobj(of-12, river-14)
Syntactic Processing: XLE Parser

- XLE parser generates dependency relations and some other important information.

<table>
<thead>
<tr>
<th>Input</th>
<th>Will-Smith was eating an apple with a fork on the bank of the river.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Dependency Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>'OBJ'(eat:8,apple:16)</td>
</tr>
<tr>
<td>'SUBJ'(eat:8,will-smith:0)</td>
</tr>
<tr>
<td>'ADJUNCT'(apple:16,on:20)</td>
</tr>
<tr>
<td>'OBJ'(on:20,bank:22)</td>
</tr>
<tr>
<td>'ADJUNCT'(apple:16,with:17)</td>
</tr>
<tr>
<td>'OBJ'(with:17,fork:19)</td>
</tr>
<tr>
<td>'ADJUNCT'(bank:22,of:23)</td>
</tr>
<tr>
<td>'OBJ'(of:23,river:25)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Important Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>'PASSIVE'(eat:8,'-')</td>
</tr>
<tr>
<td>'PROG'(eat:8,'+')</td>
</tr>
<tr>
<td>'TENSE'(eat:8,past)</td>
</tr>
<tr>
<td>'VTYPE'(eat:8,main)</td>
</tr>
</tbody>
</table>
SEMANTIC PROCESSING
Semantic Processing: Finding stems

- Wordnet is used for finding the stems of the words.

<table>
<thead>
<tr>
<th>Word</th>
<th>Stem</th>
</tr>
</thead>
<tbody>
<tr>
<td>was</td>
<td>be</td>
</tr>
<tr>
<td>eating</td>
<td>eat</td>
</tr>
<tr>
<td>apple</td>
<td>apple</td>
</tr>
<tr>
<td>fork</td>
<td>fork</td>
</tr>
<tr>
<td>bank</td>
<td>bank</td>
</tr>
<tr>
<td>river</td>
<td>river</td>
</tr>
</tbody>
</table>
Semantic Processing: WSD

<table>
<thead>
<tr>
<th>Word</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will</td>
<td>Proper</td>
</tr>
<tr>
<td>Smith</td>
<td>noun</td>
</tr>
<tr>
<td>was</td>
<td>Verb</td>
</tr>
<tr>
<td>eating</td>
<td>Verb</td>
</tr>
<tr>
<td>an</td>
<td>Determiner</td>
</tr>
<tr>
<td>apple</td>
<td>Noun</td>
</tr>
<tr>
<td>with</td>
<td>Preposition</td>
</tr>
<tr>
<td>a</td>
<td>Determiner</td>
</tr>
<tr>
<td>fork</td>
<td>Noun</td>
</tr>
<tr>
<td>on</td>
<td>Preposition</td>
</tr>
<tr>
<td>the</td>
<td>Determiner</td>
</tr>
<tr>
<td>bank</td>
<td>Noun</td>
</tr>
<tr>
<td>of</td>
<td>Preposition</td>
</tr>
<tr>
<td>the</td>
<td>Determiner</td>
</tr>
<tr>
<td>river</td>
<td>Noun</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>Synset-id</th>
<th>Sense-id</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will-Smith</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>was</td>
<td>02610777</td>
<td>be%2:42:03::</td>
</tr>
<tr>
<td>eating</td>
<td>01170802</td>
<td>eat%2:34:00::</td>
</tr>
<tr>
<td>an</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>apple</td>
<td>07755101</td>
<td>apple%1:13:00</td>
</tr>
<tr>
<td>with</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fork</td>
<td>03388794</td>
<td>fork%1:06:00::</td>
</tr>
<tr>
<td>on</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>the</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>bank</td>
<td>09236472</td>
<td>bank%1:17:01::</td>
</tr>
<tr>
<td>of</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>the</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>river</td>
<td>09434308</td>
<td>river%1:17:00::</td>
</tr>
</tbody>
</table>
Semantic Processing: Universal Word Generation

- The sense-ids are used to get the Universal Words from the Universal Word dictionary (UW++).

```
Will-Smith was eating an apple with a fork on the bank of the river.
```

<table>
<thead>
<tr>
<th>Word</th>
<th>Universal Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will-Smith</td>
<td>Will-Smith(iof&gt;PERS$ON)</td>
</tr>
<tr>
<td>eating</td>
<td>eat(icl&gt;consume, equ&gt;consumption)</td>
</tr>
<tr>
<td>apple</td>
<td>apple(icl&gt;edible fruit&gt;thing)</td>
</tr>
<tr>
<td>fork</td>
<td>fork(icl&gt;cutlery&gt;thing)</td>
</tr>
<tr>
<td>bank</td>
<td>bank(icl&gt;slope&gt;thing)</td>
</tr>
<tr>
<td>river</td>
<td>bank(icl&gt;stream&gt;thing)</td>
</tr>
</tbody>
</table>
Semantic Processing: Nouns originating from verbs

- Some nouns are originated from verbs.
- The frame *noun1 of noun2*, where *noun1* has originated from *verb1*, should generate relation *obj(verb1, noun2)* instead of any relation between *noun1* and *noun2*.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traverse the hypernym-path of the noun to its root.</td>
<td><em>removal of garbage</em> = <em>removing garbage</em></td>
</tr>
<tr>
<td>Any word on the path is “action”, then the noun has originated from a verb.</td>
<td><em>Collection of stamps</em> = <em>collecting stamps</em></td>
</tr>
<tr>
<td></td>
<td><em>Wastage of food</em> = <em>wasting food</em></td>
</tr>
</tbody>
</table>
Features of UWs

• Each UNL relation requires some properties of their UWs to be satisfied.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Word</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>agt(uw1, uw2)</td>
<td>uw1</td>
<td>Volitional verb</td>
</tr>
<tr>
<td>met(uw1, uw2)</td>
<td>uw2</td>
<td>Abstract thing</td>
</tr>
<tr>
<td>dur(uw1, uw2)</td>
<td>uw2</td>
<td>Time period</td>
</tr>
<tr>
<td>ins(uw1, uw2)</td>
<td>uw2</td>
<td>Concrete thing</td>
</tr>
</tbody>
</table>

• We capture these properties as features

• We classify the features into **Noun** and **Verb Features**
Semantic Processing: Noun Features

Algorithm

- Add word.NE_type to word.features
- Traverse the hypernym-path of the noun to its root.
- Any word on the path matches a keyword, corresponding feature generated

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add word.NE_type to word.features</td>
<td></td>
</tr>
<tr>
<td>Traverse the hypernym-path of the noun to its root.</td>
<td></td>
</tr>
<tr>
<td>Any word on the path matches a keyword, corresponding feature generated</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Words in path to hypernym-root</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>TIME</td>
</tr>
<tr>
<td>time_unit</td>
<td>TIME UNIT</td>
</tr>
<tr>
<td>time_period</td>
<td>TIME PERIOD</td>
</tr>
<tr>
<td>measure, unit</td>
<td>MEASURE</td>
</tr>
<tr>
<td>quantity, number</td>
<td>QUANTITY</td>
</tr>
<tr>
<td>person</td>
<td>PERSON</td>
</tr>
<tr>
<td>living_thing</td>
<td>ANIMATE</td>
</tr>
<tr>
<td>part</td>
<td>PART</td>
</tr>
<tr>
<td>group</td>
<td>GROUP</td>
</tr>
<tr>
<td>abstract_entity</td>
<td>ABSTRACT</td>
</tr>
<tr>
<td>physical_entity</td>
<td>CONCRETE</td>
</tr>
<tr>
<td>room, structure, facility, location, way</td>
<td>PLACE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Noun</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will-Smith</td>
<td>PERSON, ANIMATE</td>
</tr>
<tr>
<td>bank</td>
<td>PLACE</td>
</tr>
<tr>
<td>fork</td>
<td>CONCRETE</td>
</tr>
</tbody>
</table>
Types of verbs

- Verb features are based on the types of verbs

- There are three types
  - **Do**: Verbs that are performed volitionally
  - **Occur**: Verbs that are performed involuntarily or just happen without any performer
  - **Be**: Verbs which tells about existence of something or about some fact

- We classify these three into **do** and **not-do**
The sentence frames of verbs are inspected to find whether a verb is “do” type of verb or not.

“do” verbs are volitional verbs which are performed voluntarily

Heuristic:
- Something/somebody …..s something/somebody then “do” type

This heuristic fails to mark some volitional verbs as “do”, but marks all non-volitional as “not-do”.

<table>
<thead>
<tr>
<th>Word</th>
<th>Type</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>was</td>
<td>not do</td>
<td></td>
</tr>
<tr>
<td>eating</td>
<td>do</td>
<td>do</td>
</tr>
</tbody>
</table>
GENERATION OF RELATIONS AND ATTRIBUTES
Relation Generation (1/2)

• Relations handled separately

• Similar relations handled in cluster

• Rule based approach

• Rules are applied on
  – Stanford Dependency relations
  – Noun features
  – Verb features
  – POS tags
  – XLE generated information
## Relation Generation (2/2)

**Input**

\[ \text{Will-Smith was eating an apple with a fork on the bank of the river.} \]

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Relations Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{nsubj(eating, Will-Smith)}&lt;br&gt;Will-Smith.feature=ANIMATE&lt;br&gt;eat is do type and active</td>
<td>\text{agt(eat, Will Smith)}</td>
</tr>
<tr>
<td>\text{dobj(eating, apple)}&lt;br&gt;eat is active</td>
<td>\text{obj(eat, apple)}</td>
</tr>
<tr>
<td>\text{prep(eating, with), pobj(with, fork)}&lt;br&gt;fork.feature=CONCRETE</td>
<td>\text{ins(eat, fork)}</td>
</tr>
<tr>
<td>\text{prep(eating, on), pobj(on, bank)}&lt;br&gt;bank.feature=PLACE</td>
<td>\text{plc(eat, bank)}</td>
</tr>
<tr>
<td>\text{prep(bank, of), pobj(of, river)}&lt;br&gt;bank.POS = noun&lt;br&gt;river.feature is not PERSON</td>
<td>\text{mod(bank, river)}</td>
</tr>
</tbody>
</table>
Attribute Generation (1/2)

- Attributes handled in clusters
- Attributes are generated by rules
- Rules are applied on
  - Stanford dependency relations
  - XLE information
  - POS tags
Attribute Generation (2/2)

Input

Will-Smith was eating an apple with a fork on the bank of the river.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Attributes Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>eat is verb of main clause</td>
<td>eat</td>
</tr>
<tr>
<td>VTYPE(eat, main)</td>
<td>@entry</td>
</tr>
<tr>
<td>PROG(eat,+)</td>
<td>@progress</td>
</tr>
<tr>
<td>TENSE(eat,past)</td>
<td>@past</td>
</tr>
<tr>
<td>det(apple, an)</td>
<td>apple@indef</td>
</tr>
<tr>
<td>det(fork, a)</td>
<td>fork@indef</td>
</tr>
<tr>
<td>det(bank, the)</td>
<td>bank@def</td>
</tr>
<tr>
<td>det(river, the)</td>
<td>river@def</td>
</tr>
</tbody>
</table>
EVALUATION OF OUR SYSTEM
Results

• Evaluated on EOLSS corpus
  – 7000 sentences from EOLSS corpus

• Evaluated on 4 scenarios
  – Scenario1: Previous system (tightly coupled with XLE parser)
  – Scenario2: Current system (tightly coupled with Stanford parser)
  – Scenario3: Scenario2 – (Simplifier + Merger)
  – Scenario4: Scenario2 – XLE parser

• Scenario3 lets us know the impact of Simplifier and Merger

• Scenario4 lets us know the impact of XLE parser
Results: Metrics

- \( t_{\text{gen}} = \text{relation}_{\text{gen}} (UW_{1\text{gen}}, UW_{2\text{gen}}) \)
- \( t_{\text{gold}} = \text{relation}_{\text{gold}} (UW_{1\text{gold}}, UW_{2\text{gold}}) \)

- \( t_{\text{gen}} = t_{\text{gold}} \) if
  - \( \text{relation}_{\text{gen}} = \text{relation}_{\text{gold}} \)
  - \( UW_{1\text{gen}} = UW_{1\text{gold}} \)
  - \( UW_{2\text{gen}} = UW_{2\text{gold}} \)

- If \( UW_{\text{gen}} = UW_{\text{gold}} \), let \( a_{\text{match}}(UW_{\text{gen}}, UW_{\text{gold}}) = \{\# \text{ attributes matched in } \text{UW}_{\text{gen}} \text{ and } \text{UW}_{\text{gold}}\} \)

- \( \text{count}(UW_x) = \{\# \text{ attributes in } UW_x\} \)
Results: Relation-wise accuracy

\[
\begin{align*}
\text{Precision } p_{relation} &= \frac{\#(t_{gen,relation} = t_{gold,relation})}{\#t_{gen,relation}} \\
\text{Recall } r_{relation} &= \frac{\#(t_{gen,relation} = t_{gold,relation})}{\#t_{gold,relation}} \\
\text{F-score } f_{relation} &= \frac{2 \times p_{relation} \times r_{relation}}{p_{relation} + r_{relation}}
\end{align*}
\]
Results: Overall accuracy

Precision $p_{overall} = \frac{\#(t_{gen} = t_{gold})}{\#t_{gen}}$

Recall $r_{overall} = \frac{\#(t_{gen} = t_{gold})}{\#t_{gold}}$

F-score $f_{overall} = \frac{2 \times p_{overall} \times r_{overall}}{p_{overall} + r_{overall}}$

Attribute accuracy $= \sum_{UW_{gold}=UW_{gen}, UW_{gold} \in GEN, UW_{gen} \in GOLD} a_{match}(UW_{gold}, UW_{gen})$

$= \frac{\sum_{UW_{gold}=UW_{gen}, count(UW_{gold})} a_{match}(UW_{gold}, UW_{gen})}{\sum_{UW_{gold}=UW_{gen}, count(UW_{gold})} count(UW_{gold})}$
Results: Relation
Results: Relation-wise: Precision
Results: Relation-wise: Recall
Results: Relation-wise: F-score
Results: Attribute

![Attribute Accuracy Chart]

- Scenario1: Accuracy 0.37
- Scenario2: Accuracy 0.39
- Scenario3: Accuracy 0.45
- Scenario4: Accuracy 0.42
Results: Time-taken

- Time taken for the systems to evaluate 7000 sentences from EOLSS corpus.
ERROR ANALYSIS
Wrong Relations Generated

False positive

- obj
- mod
- aoj
- man
- and
- agt
- plc
- qua
- pur
- gol
- tim
- or
- overall

False positive
Missed Relations

False negative

[Bar chart showing missed relations for different categories like obj, mod, aoj, man, and, agt, plc, qua, pur, gol, tim, or, and overall. The overall category has the highest number of false negatives.]
Case Study (1/3)

• Conjunctive relations
  – Wrong apposition detection
    • James loves eating red water-melon, apples peeled nicely and bananas.

```
 amod(water-melon-5, red-4)
 appos(water-melon-5, apples-7)*
 partmod(apples-7, peeled-8)
 advmod(peeled-8, nicely-9)
 cc(nicely-9, and-10)
 conj(nicely-9, bananas-11)
```
conjunctive relations

- Non-uniform relation generation

- George, the president of the football club, James, the coach of the club and the players went to watch their rival's game.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Argument 1</th>
<th>Argument 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>appos</td>
<td>George-1</td>
<td>president-4</td>
</tr>
<tr>
<td>conj</td>
<td>club-8</td>
<td>James-10</td>
</tr>
<tr>
<td>conj</td>
<td>club-8</td>
<td>coach-13</td>
</tr>
<tr>
<td>conj</td>
<td>club-8</td>
<td>players-19</td>
</tr>
<tr>
<td>cc</td>
<td>club-8</td>
<td>and-17</td>
</tr>
</tbody>
</table>
Case Study (3/3)

- **Low Precision frequent relations**
  - mod and qua
  - More general rules
  - `agt-aoj`: non-uniformity in corpus
  - `Gol-obj`: Multiple relation generation possibility

The former, however, represents the toxic concentration in the water, while the latter represents the dosage to the body of the test organisms.

<table>
<thead>
<tr>
<th>Corpus Relations</th>
<th>Generated Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>agt(represent, latter)</code></td>
<td><code>aoj(represent, latter)</code></td>
</tr>
<tr>
<td><code>aoj(represent, former)</code></td>
<td><code>aoj(represent, former)</code></td>
</tr>
<tr>
<td><code>obj(dosage, body)</code></td>
<td><code>gol(dosage, body)</code></td>
</tr>
</tbody>
</table>
Hindi Generation from Interlingua (UNL)

(Joint work with S. Singh, M. Dalal, V. Vachhani, Om Damani
MT Summit 2007)
Deconversion = Transfer + Generation
Step-through the Deconverter

<table>
<thead>
<tr>
<th>Module</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNL Expression</td>
<td><code>obj(contact(…</code></td>
</tr>
<tr>
<td>Lexeme Selection</td>
<td>संपक किसान यह आप क्षेत्र तालुक मंचर खटाव contact farmer this you region taluka manchar khatav</td>
</tr>
<tr>
<td>Case Identification</td>
<td>संपक किसान यह आप क्षेत्र तालुक मंचर खटाव contact farmer this you region taluka manchar khatav</td>
</tr>
<tr>
<td>Morphology Generation</td>
<td>संपक कीजिए किसानों इस आप क्षेत्र contact .@imperative farmer.@@pl this you region taluka मंचर खटाव taluka manchar Khatav</td>
</tr>
<tr>
<td>Function Word Insertion</td>
<td>संपक कीजिए किसानों को इसके लिए आप क्षेत्र contact farmers this for you region या तालुक के मंचर खटाव</td>
</tr>
<tr>
<td>Linearization</td>
<td>इसके लिए आप मंचर क्षेत्र या खटाव This for you manchar region or khatav</td>
</tr>
</tbody>
</table>
Lexeme Selection

[संपर्क]{}"contact(icl>communicate(agt>person,obj>person))“ (V, VOA, VOA-ACT, VOA-COMM, VLTN, TMP, CJNCT, N-V, link, Va)
[पहचान कार्यक्षित]{}"contact(icl>representative)“
(N, ANIMT, FAUNA, MML, PRSN, Na)

Lexical Choice is unambiguous
Case Marking

- Depends on UNL Relation and the properties of the nodes
- Case get transferred from head to modifiers
The boy saw me.
लड़के ने मुझे देखा।
Boys saw me.
लड़कों ने मुझे देखा।
The King saw me.
राजा ने मुझे देखा।
Kings saw me.
राजाों ने मुझे देखा।

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Attribute values</th>
</tr>
</thead>
<tbody>
<tr>
<td>uoM</td>
<td>@N, @NU, @M, @pl, @oblique</td>
</tr>
<tr>
<td>U</td>
<td>@N, @NU, @M, @sg, @oblique</td>
</tr>
<tr>
<td>I</td>
<td>@N, @NI, @F, @sg, @oblique</td>
</tr>
<tr>
<td>iyoM</td>
<td>@N, @NI, @F, @pl, @oblique</td>
</tr>
<tr>
<td>oM</td>
<td>@N, @NA, @NOTCH, @F, @pl, @oblique</td>
</tr>
</tbody>
</table>
## Verb Morphology

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Tense</th>
<th>Aspect</th>
<th>Mood</th>
<th>N</th>
<th>Gen</th>
<th>P</th>
<th>V</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>-e rahaa thaa</td>
<td>@past</td>
<td>@progress</td>
<td>-</td>
<td>@sg</td>
<td>@male</td>
<td>3\textsuperscript{rd}</td>
<td>e</td>
<td></td>
</tr>
<tr>
<td>-taa hai</td>
<td>@present</td>
<td>@custom</td>
<td>-</td>
<td>@sg</td>
<td>@male</td>
<td>3\textsuperscript{rd}</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>-iyaa thaa</td>
<td>@past</td>
<td>@complete</td>
<td>-</td>
<td>@sg</td>
<td>@male</td>
<td>3\textsuperscript{rd}</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>saktii hain</td>
<td>@present</td>
<td>-</td>
<td>@ability</td>
<td>@pl</td>
<td>@female</td>
<td>3\textsuperscript{rd}</td>
<td>A</td>
<td></td>
</tr>
</tbody>
</table>
After Morphology Generation

contact संपर्क
to farmer किसान
to you आप

contact संपर्क कीजिए
to farmer किसानी
to you आप

this यह
this इस
Function Word Insertion

संपर्क कीजिए किसानों यह आप क्षेत्र तालुके मंचर खटाव
संपर्क कीजिए किसानों को इसके लिए आप क्षेत्र या तालुके के मंचर खटाव

<table>
<thead>
<tr>
<th>Rel</th>
<th>Par+</th>
<th>Par-</th>
<th>Chi+</th>
<th>Chi-</th>
<th>Ch/FW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obj</td>
<td>V</td>
<td>VINT</td>
<td>N#ANIMT</td>
<td>@topic</td>
<td>को</td>
</tr>
</tbody>
</table>

obj

contact संपर्क कीजिए

agt

farmer किसानों को

pur

this इसके लिए

you आप
Linearization

इस आप मंचर क्षेत्र
This you manchar region

खटाव तालुके किसानों
khatav taluka farmers

संपर्क
contact
Syntax Planning: Assumptions

- The relative word order of a UNL relation’s relata does not depend on:
  - Semantic Independence: the semantic properties of the relata.
  - Context Independence: the rest of the expression.
- The relative word order of various relations sharing a relatum does not depend on
  - Local Ordering: the rest of the expression.
Syntax Planning: Strategy

- Divide a nodes relations in
  Before_set = \{obj,pur,agt\}
  After_set = {}  

Topo Sort each group:
- pur agt obj
- this you farmer

Final order: this you farmer contact
## Syntax Planning Algo

<table>
<thead>
<tr>
<th>Stack</th>
<th>Before Current</th>
<th>After Current</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>region farmer contact</td>
<td></td>
<td></td>
<td>this you</td>
</tr>
<tr>
<td>manchar taluka</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>manchar region taluka farmer contact</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>region taluka farmer Contact</td>
<td></td>
<td></td>
<td>this you manchar</td>
</tr>
<tr>
<td>taluka farmer contact</td>
<td></td>
<td></td>
<td>this you manchar region</td>
</tr>
</tbody>
</table>

![Diagram](image-url)
## All Together (UNL -> Hindi)

<table>
<thead>
<tr>
<th>Module</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UNL Expression</strong></td>
<td>obj(contact(..</td>
</tr>
<tr>
<td><strong>Lexeme Selection</strong></td>
<td>संपर्क किसान् यह आप क्षेत्र तालुक मंचर खटाव contact farmer this you region taluka manchar khatav</td>
</tr>
<tr>
<td><strong>Case Identification</strong></td>
<td>संपर्क किसान् यह आप क्षेत्र तालुक मंचर खटाव contact farmer this you region taluka manchar khatav</td>
</tr>
<tr>
<td><strong>Morphology Generation</strong></td>
<td>संपर्क कीजिए किसानों यह आप क्षेत्र contact .@imperative farmer.@pl this you region taluka मंचर खटाव taluka manchar Khatav</td>
</tr>
<tr>
<td><strong>Function Word Insertion</strong></td>
<td>संपर्क कीजिए किसानों को इसके लिए आप क्षेत्र contact farmers this for you region क्षेत्र या तालुक के मंचर खटाव or taluka of Manchar Khatav</td>
</tr>
<tr>
<td><strong>Syntax Linearization</strong></td>
<td>इसके लिए आप मंचर क्षेत्र या खटाव This for you manchar region or khatav तालुक के किसानों को संपर्क कीजिए</td>
</tr>
</tbody>
</table>
How to Evaluate UNL Deconversion

- UNL -> Hindi

- Reference Translation needs to be generated from UNL
  - Needs expertise

- Compromise: Generate reference translation from original English sentence from which UNL was generated
  - Works if you assume that UNL generation was perfect

- Note that fidelity is not an issue
**Manual Evaluation Guidelines**

**Fluency** of the given translation is:
(4) Perfect: Good grammar  
(3) Fair: Easy-to-understand but flawed grammar  
(2) Acceptable: Broken - understandable with effort  
(1) Nonsense: Incomprehensible

**Adequacy**: How much meaning of the reference sentence is conveyed in the translation:  
(4) All: No loss of meaning  
(3) Most: Most of the meaning is conveyed  
(2) Some: Some of the meaning is conveyed  
(1) None: Hardly any meaning is conveyed
<table>
<thead>
<tr>
<th>Hindi Output</th>
<th>Fluency</th>
<th>Adequacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>कवक के कारण आम की नाजुक पत्तियां यदि झुलसे रहे हैं 0.5 प्रतिशत का बोर्ड मिश्रण 10 लीटर पानी के साथ तो छिड़का जाना चाहिए</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>परीक्षण के अनुसार नियमित रूप से फलों की अच्छी वृद्धि के लिए खादों की खुराक के दी जानी चाहिए</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>हमें मेथी का फसल के बाद बोता है मेथी और धनिया फसल को अच्छी बढ़ने या अच्छे बढ़ने बता</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>जीवाणिक संक्रमण से इसकी जड़ें प्रभावित होती हैं</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>इमु का पक्षी रेटाइट का परिवार को संबंधित होता है और थोड़ा यह शुतुरमुग के साथ समान दिखती हैं</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>Fluency</th>
<th>Adequacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometric Average</td>
<td>0.34</td>
<td>2.54</td>
<td>2.84</td>
</tr>
<tr>
<td>Arithmetic Average</td>
<td>0.41</td>
<td>2.71</td>
<td>3.00</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.25</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>Pearson Cor. BLEU</td>
<td>1.00</td>
<td>0.59</td>
<td>0.50</td>
</tr>
<tr>
<td>Pearson Cor. Fluency</td>
<td>0.59</td>
<td>1.00</td>
<td>0.68</td>
</tr>
</tbody>
</table>

- Good Correlation between Fluency and BLEU
- Strong Correlation between Fluency and Adequacy
- Can do large scale evaluation using Fluency alone

Caution: Domain diversity, Speaker diversity
Summary: interlingua based MT

- English to Hindi
- Rule governed
- High level of linguistic expertise needed
- Takes a long time to build (since 1996)
- But produces great insight, resources and tools