A resource-light approach to learning verb valencies

Alex Rudnick
alexr@cs.indiana.edu

MTML 26 January 2011

spilling the beans

- Ongoing work: trying to learn which arguments go with which verbs in a very restricted resource situation.

- For Quechua, all we have is some text and a morphological analyzer.

- Plan: Use another morphologically rich language with more resources for comparison. We have a treebank for Arabic!

- Results: to learn verb subcats, we probably really want at least a chunker, and definitely a morphological disambiguator.

outline

- **context: what we're trying to do, and why**
- about Quechua
- attempts to learn about Arabic verbs, for comparison
- trying it out for Quechua
- the future

goals!

- Aiming for high-quality translation over limited domains

- resource-scarce languages: Quechua, Amharic, Tigrinya, Oromo...

- taking linguistic knowledge seriously

- synchronous, multi-layer dependency grammar, with a lexicon

XDG

- eXtensible Dependency Grammar.

- Originally built by Ralph Debusmann, using Mozart/Oz.

- Constraint-based dependency grammar framework; solutions are multigraphs. Constraints can be either within a layer or across layers of a solution.

- Originally used for parsing; generation has been done too.

- We've got a version in Python

Synchronous XDG

- Mike Gasser's talk last week at FreeRBMT in Barcelona: "Toward Synchronous Extensible Dependency Grammar"

- We (Mike) have really basic English-Amharic translation going for sentences about water.

"Esther filtered water" produces both:

ውሀ አስቴር አጠለለች . (OSV)

አስቴር ውሀ አጠለለች . (SOV)

different verbs take different arguments

"What did you do yesterday?"
* "I put, mostly."
* "I put the potato."
"I put the potato on the table."

"I ate the potato."
"I ate already."

"I believe that he is tall."
"I consider him tall."
* "I consider that he is tall."

verb subcategories are an issue for us

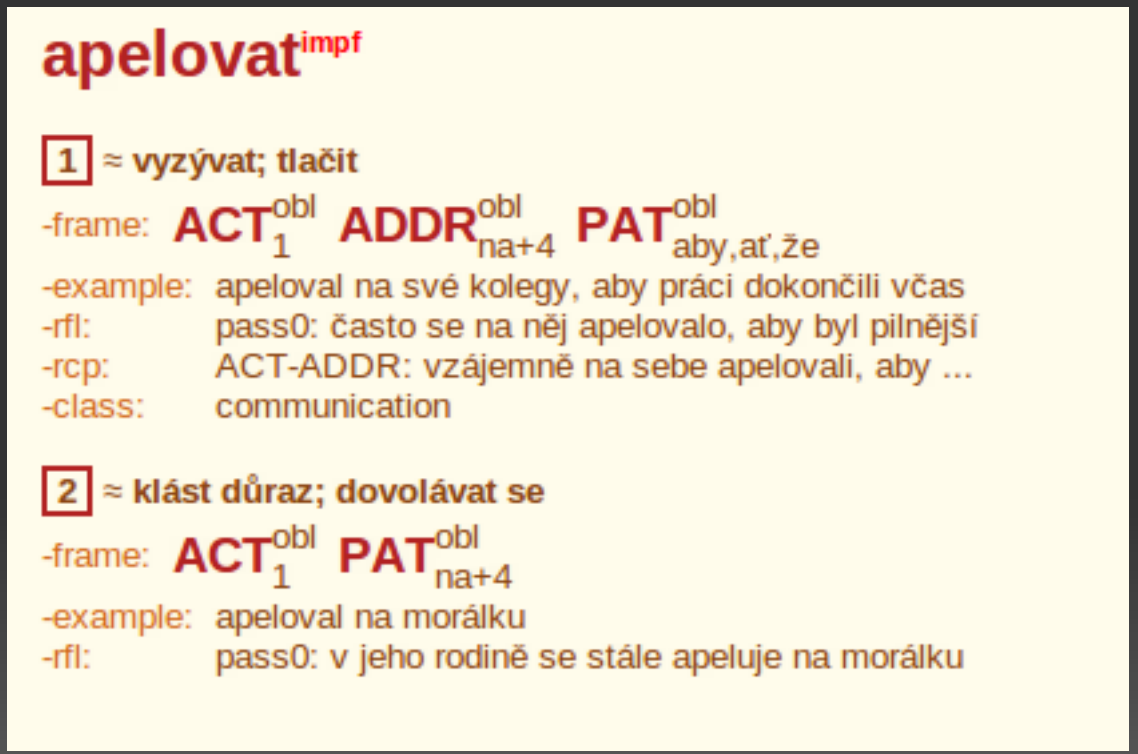
- we have an explicit lexicon

- a solution writes down the dependents of a verb

- system is currently unweighted, but we'd like to put weights on rules for ranking parses/translations and to be more robust.

what we'd really like

VALLEX for Czech, although it only has so many verbs, and it's unweighted...



valency vs. subcat

I shouldn't use them interchangeably: I really just mean subcategory: which arguments should the verb look for?

In Functional Generative Description (FGD), valency refers to a tectogrammatical notion; arguments might be known to the speaker but not expressed.

Bojar points out that this can't really be observed from the text alone...

outline

- context: what we're trying to do, and why
- **about Quechua**
- attempts to learn about Arabic verbs, for comparison
- trying it out for Quechua
- the future

Quechua

About 10 million native speakers, more than Swedish.
Biggest Native American language; really a dialect continuum.
Spoken in Peru, Ecuador, Bolivia -- language of the Inca empire.

Quechua social context

- Indigenous language

- many speakers are bilingual with Spanish

- Revitalization effort underway, with bilingual education programs

- Academia Mayor de la Lengua Quechua: wants to establish the Cuzco dialect as "official" and "pure".

- cool hats in the Andes

Quechua morphology

- Agglutinative language, quite regular

wasi	"house" (as subject of sentence)
wasi+rayku	"because of the house" (causative)
wasi+pi+wan	"including the house" (locative, instrumental)

- Also, verbs can have evidentiality! (how to generate that? ...)

outline

- context: what we're trying to do, and why

- about Quechua

- **attempts to learn about Arabic verbs, for comparison**

- trying it out for Quechua

- the future

more resources for Arabic

Using the Penn Arabic Treebank, we can find the verbs and their arguments easily. Better than having a deep parser: now we have presumably-correct parses!

PATB comes with the same sentences, un-parsed and in Unicode; treebanks are transliterated.

morphological analysis for Arabic

- Buckwalter morphological analyzer is often used for Arabic

- Using Pierrick Brihaye's port of Buckwalter, AraMorph (Java, GPL, handles unicode text)

evaluation

- For Arabic, compare precision/recall on subcat frames using the treebank.

- Find verbs that we ignore a lot: why do we ignore them? (auxiliary-type verbs, or ones that go with embedded clauses?)

heuristics for picking sentences

- Like Przepiorkowski (2009), dropping sentences with more than one verb, to avoid deciding attachment issues.

- Bojar (2003) does something similar, but can often chunk subordinate clauses.

Helpfully, case is often marked on nouns in Quechua and Arabic -- so we should be able to find the arguments we want, right?

(what if they're dropped? ...)

heuristics for picking sentences

- But this leaves only 24% of the sentences in the AFP corpus: 1405 out of 5845.

- In the 1405 1-verb sentences, we see 506 different verbs (inflected); not very good use of the data!

- Worse, many of the verbs just won't be represented -- what about ones that go with a clause? (think, believe, request...)

Arabic morphological ambiguity

- Very rarely get a unique morphological analysis...

- Within the Penn ATB, there are about 7.5 analyses per word. (max: 86, stddev: 8.4)

86 analyses: وَاحِد: "one", "and scrutinize", "and sharpen", "and be furious".

coverage

- If our goal is broad coverage, how many verbs will you observe in the treebank?

"The total number of Czech verbs is difficult to estimate, but is expected to be around 40,000. The PDT [Prague Dependency Treebank] covers only 5,407 of these verbs..." (Bojar 2003)

PATB is similar: part 1 only has about 26 thousand types at all. (out of about 166 thousand tokens)

coverage

On its own, this isn't a problem -- it's promising!

To improve coverage, we could just feed in more un-annotated text; there's basically an endless supply of Arabic text available.

Less plausible for Quechua...

outline

- context: what we're trying to do, and why

- about Quechua

- attempts to learn about Arabic verbs, for comparison

- **trying it out for Quechua**

- the future

Quechua resources

For Quechua, we have a small corpus from CMU's AVENUE project:
- elicitation sentences: bitext!
- some UN documents, a few stories
- a bunch of scanned book pages in various formats...

Also, several dictionaries are online!
http://www.runasimipi.org/blanco.php?file=diccionarios
(we should parse these)

Quechua resources

We also have AntiMorfo, a morphological analyzer/generator that Mike Gasser put together.

- "Anti" is the word for "Andes" in Quechua.

- it doesn't do the opposite of morphology

morphological ambiguity in Quechua

- Quechua has a similar issue, but not nearly so bad. Probably because of vowels!

mean: 1.7 analyses
stddev: 1.3
max: 10

- We can't currently tell whether a word is definitely a verb.
waqaychu: verb? noun? infinitive?
(waqa is in vocabulary as a verb root, waqay as noun root)

- So we're going to need (at least) a POS tagger...

outline

- context: what we're trying to do, and why

- about Quechua

- attempts to learn about Arabic verbs, for comparison

- trying it out for Quechua

- **the future**

finding a resource sweet spot...

- It's almost definitely not enough to just use a morphological analyzer and example text.

- Przepiorkowski uses a chunker for Polish; Bojar has finite-state rules for finding clauses and coordination in Czech.

- There's MADA+TOKAN for morphological disambiguation in Arabic; maybe that would let us use text from the wild

- But only for Arabic.

- For Quechua, we should probably parse dictionaries and maybe work on a chunker. Maybe children's books?

thoughts? questions?

Thanks!

(not looking for a postdoc yet; it'll be a few years.)