Corpus Linguistics: Solutions for Annotation

Karën Fort
karen.fort@inist.fr

December 12, 2011

Introduction

Solutions

Solutions

- Annotation Tools
- Tag Dictionaries / Pre-annotation / Active Learning
- Crowdsourcing (AMT and serious games)
- Training / Documentation / Methodology

Why?

Why using tools?

- To ease the editing of annotations, in particular in the case of relations
- To limit the number of items to keep in mind
  [Dandapat et al., 2009]
- To constraint the annotation, therefore limiting the errors
  [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]
- To hide a layer when annotating another one
  [Widlöcher and Mathet, 2009]
- To ease the access to the context, even large
  [Widlöcher and Mathet, 2009]
- To keep track of the discussions between annotators
  [Lortal et al., 2006] or of the errors and their corrections
  [de la Clergerie, 2008]

Why using tools?

- To ease the editing of annotations, in particular in the case of relations
- To limit the number of items to keep in mind [Dandapat et al., 2009]
- To constraint the annotation, therefore limiting the errors [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]
- To hide a layer when annotating another one [Widlöcher and Mathet, 2009]
- To ease the access to the context, even large [Widlöcher and Mathet, 2009]
- To keep track of the discussions between annotators [Lortal et al., 2006] or of the errors and their corrections [de la Clergerie, 2008]

Why using tools?

- To ease the editing of annotations, in particular in the case of relations
- To limit the number of items to keep in mind
  [Dandapat et al., 2009]
- To constraint the annotation, therefore limiting the errors
  [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]
- To hide a layer when annotating another one
  [Widlöcher and Mathet, 2009]
- To ease the access to the context, even large
  [Widlöcher and Mathet, 2009]
- To keep track of the discussions between annotators
  [Lortal et al., 2006] or of the errors and their corrections
  [de la Clergerie, 2008]

Why using tools?

- To ease the editing of annotations, in particular in the case of relations
- To limit the number of items to keep in mind
  [Dandapat et al., 2009]
- To constraint the annotation, therefore limiting the errors
  [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]
- To hide a layer when annotating another one
  [Widlöcher and Mathet, 2009]
- To ease the access to the context, even large
  [Widlöcher and Mathet, 2009]
- To keep track of the discussions between annotators
  [Lortal et al., 2006] or of the errors and their corrections
  [de la Clergerie, 2008]

Why using tools?

- To ease the editing of annotations, in particular in the case of relations
- To limit the number of items to keep in mind
  [Dandapat et al., 2009]
- To constraint the annotation, therefore limiting the errors
  [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]
- To hide a layer when annotating another one
  [Widlöcher and Mathet, 2009]
- To ease the access to the context, even large
  [Widlöcher and Mathet, 2009]
- To keep track of the discussions between annotators
  [Lortal et al., 2006] or of the errors and their corrections
  [de la Clergerie, 2008]

Why using tools?

- To ease the editing of annotations, in particular in the case of relations
- To limit the number of items to keep in mind
  [Dandapat et al., 2009]
- To constraint the annotation, therefore limiting the errors
  [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]
- To hide a layer when annotating another one
  [Widlöcher and Mathet, 2009]
- To ease the access to the context, even large
  [Widlöcher and Mathet, 2009]
- To keep track of the discussions between annotators
  [Lortal et al., 2006] or of the errors and their corrections
  [de la Clergerie, 2008]

Existing tools

+/- Glozz, GATE, but also MMAX2, Knowtator, Cadixe, Callisto, etc.

++ gain in time and quality

$\Rightarrow$ (too) many tools, for schemes or for specific campaign, not for annotators!

Existing tools

+/- Glozz, GATE, but also MMAX2, Knowtator, Cadixe, Callisto, etc.

++ gain in time and quality

$\Rightarrow$ (too) many tools, for schemes or for specific campaign, not for annotators!

Can XML editors be considered as annotation tools?

Tag Dictionaries

Allow to:

1. store the categories attached by annotators to one token
2. propose those categories when the same token is met

$\Rightarrow$ Very simple and quite effective (see [Carmen et al., 2010]), but the more is annotated, the more effective the method is.

Correcting automatic pre-annotations

++ Significative gain in time and quality, at least for POS tagging and bracketing (Penn Treebank [Marcus et al., 1993], Hindi and Bangla POS tagging [Dandapat et al., 2009], English POS tagging [Fort and Sagot, 2010])

- Biases not always taken into account: is it the same to pre-annotate NEs and gene renaming?
- also time consuming if system is too bad (to be defined)

Particular case: Active Learning

- Not all the annotations are necessary to train a tool $\Rightarrow$ detect annotations that are really useful to improve the final results
- Pre-annotate a corpus automatically, then ask annotators to correct, then re-annotate, etc.

$\Rightarrow$ iterative

\+ allow to gain time

\- but time consuming if system is too bad (to be defined)

- on Ritel project (Human Machine Oral Dialog): above 30% of errors, it was quickier for transcriber to do it from scratch than to correct transcription

Pre-annotations issues

- Either humans concentrate on what was pre-annotated, correct pre-annotations, **but** do not see what is missing
- or they concentrate on what is missing **but** do not correct pre-annotations.
- impossible for some types of annotation due to the lack of good quality tools (like co-reference resolvers)

Training and Documentation

A good training of the annotators is the best solution for a fast, better quality annotation [Dandapat et al., 2009].

This should be associated with an adapted documentation with:

- a clear definition of the **application**
- a clear and detailed definition of the **categories** (always possible or even desirable?)
- meaningful **examples**
- **ambiguous categories** presented in parallel, like in the PTB documentation (see it here: ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz)

Training and Documentation

A good training of the annotators is the best solution for a fast, better quality annotation [Dandapat et al., 2009].

This should be associated with an adapted documentation with:

- a clear definition of the **application**
- a clear and detailed definition of the **categories** (always possible or even desirable?)
- meaningful **examples**
- **ambiguous categories** presented in parallel, like in the PTB documentation (see it here: ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz)

Keep in mind that annotators are **at the very heart** of the annotation campaign!

Methodology

- Compute inter-annotator agreement at the very beginning of the campaign, then update the Annotation Guide [Bonneau-Maynard et al., 2005].
- Compute intra-annotator agreement as the annotation goes, to check that annotators are coherent with themselves [Gut and Bayerl, 2004].
- This can go as far as Agile Annotation [Voormann and Gut, 2008, Alex et al., 2010], implying several iterations

Crowdsourcing: Definition

Crowdsourcing *is the act of outsourcing tasks, traditionally performed by an employee or contractor, to an undefined, large group of people or community (a crowd), through an open call.*

The term "crowdsourcing" is a portmanteau of "crowd" and "outsourcing", first coined by Jeff Howe in a June 2006 Wired magazine article "The Rise of Crowdsourcing". Howe explains that because technological advances have allowed for cheap consumer electronics, the gap between professionals and amateurs has been diminished. Companies are then able to take advantage of the talent of the public, and Howe states that "It's not outsourcing; it's crowdsourcing."

(Wikipedia, consulted on the 2nd of Dec., 2010)

Different types of crowdsourcing

Developed with Web 2.0:

- crowdvoting: using social networks to vote on an issue, a product, etc (social bookmarking)
- crowdcreation: idea competitions
- crowdwisdom: answering questions (Yahoo! questions)
- crowfunding (for art projects, political campaigns, etc)

... through social networks, "serious" games and microworking.

Different types of crowdsourcing

Developed with Web 2.0:

- crowdvoting: using social networks to vote on an issue, a product, etc (social bookmarking)
- crowdcreation: idea competitions
- crowdwisdom: answering questions (Yahoo! questions)
- crowfunding (for art projects, political campaigns, etc)

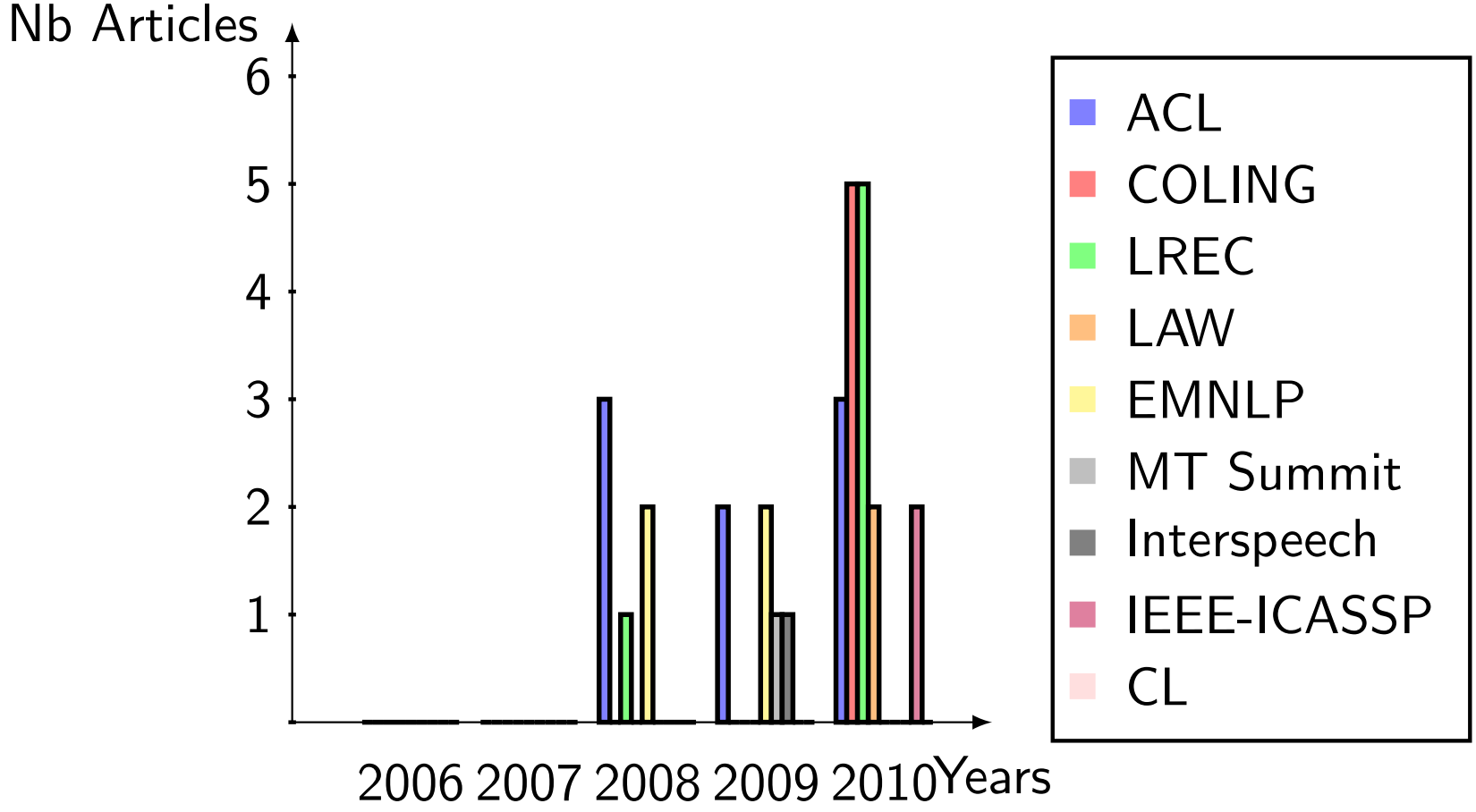
... through ~~social networks~~, **"serious" games** and **microworking**.

Crowdsourcing: Serious Games

- ESP game: 13,500 users labelled 1.3M images in 3 months! [von Ahn, 2006]
- JeuxDeMots [Lafourcade, 2007]
- PhraseDetectives [Chamberlain et al., 2008]

Crowdsourcing: Microworking

Amazon Mechanical Turk (AMT): ACL Anthology (Nov. 5, 2010), 86 art. (incl. NAACL-HLT 2010 Workshop)



Evolution of MTurk usage in NLP [Fort et al., 2011]

MTurk: Gold Mine or Coal Mine? [Fort et al., 2011]

Presentation, LTC 2011 [Adda et al., 2011]

Crowdsourcing: pros and cons

+ using the users' work through Web collaboration (access to more people)
  - who's working? (native language? Education?)
+ cheap (if not free)
  - from games to hobby to... sweatshop!
+ quick
+ good quality [Snow et al., 2008]...
  - ... if annotation is easy!
  - and if people do not cheat!

- solutions
- pros and cons


Adda, G., Sagot, B., Karën Fort, and Mariani, J. (2011).
Crowdsourcing for language resource development: Critical analysis of amazon mechanical turk overpowering use.
In Proc. of the Language and Technology Conference, Poznań, Pologne.


Alex, B., Grover, C., Shen, R., and Kabadjov, M. (2010).
Agile corpus annotation in practice: An overview of manual and automatic annotation of cvs.
In Proceedings of the Fourth Linguistic Annotation Workshop, pages 29–37, Uppsala, Sweden. Association for Computational Linguistics.


Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., and Mostefa, D. (2005).
Semantic Annotation of the French Media Dialog Corpus.
In InterSpeech, Lisboa, Portugal.


Carmen, M., Felt, P., Haertel, R., Lonsdale, D., McClanahan, P., Merkling, O., Ringger, E., and Seppi, K. (2010).

Tag dictionaries accelerate manual annotation.
In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), La Valette, Malte. European Language Resources Association (ELRA).


Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008).
Phrase Detectives: a Web-based Collaborative Annotation Game.
In Proceedings of the International Conference on Semantic Systems (I-Semantics'08), Graz.


Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009).
Complex Linguistic Annotation - No Easy Way Out! A Case from Bangla and Hindi POS Labeling Tasks.
In Proceedings of the third ACL Linguistic Annotation Workshop, Singapour.


de la Clergerie, E. V. (2008).
A Collaborative Infrastructure for Handling Syntactic Annotations.

In First International Workshop on Automated Syntactic Annotations for interoperable Language Resources, Hong-Kong.


Fort, K., Adda, G., and Cohen, K. B. (2011).
Amazon mechanical turk: Gold mine or coal mine?
Computational Linguistics (editorial), 37(2).


Fort, K. and Sagot, B. (2010).
Influence of Pre-annotation on POS-tagged Corpus Development.
In Proc. of the Fourth ACL Linguistic Annotation Workshop, Uppsala, Suède.


Gut, U. and Bayerl, P. S. (2004).
Measuring the Reliability of Manual Annotations of Speech Corpora.
In Proceedings of Speech Prosody, pages 565–568, Nara, Japan.


Lafourcade, M. (2007).

Making people play for lexical acquisition.
In Proc. SNLP 2007, 7th Symposium on Natural Language Processing, Pattaya, Thailande.


Lortal, G., Todirascu-Courtier, A., and Lewkowicz, M. (2006).
Soutenir la coopération par l'indexation semi-automatique d'annotations.
In Actes de la Semaine de la Connaissance 2006, Nantes, France.


Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993).
Building a large annotated corpus of english : The penn treebank.
Computational Linguistics, 19(2):313–330.


Mikulová, M. and Štěpánek, J. (2009).
Annotation quality checking and its implications for Design of treebank (in building the prague czech-english Dependency treebank).

In Proceedings of the Eight International Workshop on Treebanks and Linguistic Theories, volume 4-5, Milan, Italie.


Snow, R., O'Connor, B., Jurafsky, D., and Ng., A. Y. (2008).
Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks.
In Proceedings of EMNLP 2008, pages 254–263.


von Ahn, L. (2006).
Games with a purpose.
IEEE Computer Magazine, pages 96–98.


Voormann, H. and Gut, U. (2008).
Agile corpus creation.
Corpus Linguistics and Linguistic Theory, 4(2):235–251.


Widlöcher, A. and Mathet, Y. (2009).
La plate-forme glozz : environnement d'annotation et d'exploration de corpus.
In Actes de Traitement Automatique des Langues 2009 (TALN 2009), Senlis, France.