Local Modifications and Paraphrases in Wikipedia's Revision History

Camille Dutrey, Houda Bouamor, **Delphine Bernhard** and Aurélien Max

LIMSI-CNRS & Univ. Paris Sud, Orsay, France

CBA 2010, Corpus-Based Approaches to Paraphrasing and Nominalization

Plan

Plan

Wikipedia, the free encyclopedia

Project and Objectives

- Universal, multilingual and free encyclopedia
- Freely accessible and reusable content

Global statistics, August 4. 2010

- 270 language versions
- 15,301,474 articles

Statistics for the French version, September 15. 2010

- 993,967 articles
- 60,592,371 modifications

Wikipedia's Revision History

Revision history of Paraphrase

From Wikipedia, the free encyclopedia
View logs for this page

Browse history
From year (and earlier): From month (and earlier): all Tag filter: Deleted only Go

For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary.
External tools: Revision history statistics · Revision history search · Number of watchers · Page view statistics

(cur) = difference from current version, (prev) = difference from preceding version, **m** = minor edit, → = section edit, ← = automatic edit summary
(latest | earliest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

- (cur | prev) 22:18, 18 November 2010 ClueBot NG (talk | contribs) **m** (3,508 bytes) *(Reverting possible vandalism by 98.228.34.169 to version by 96.51.77.240. Questions, comments, complaints -> BRFA Thanks, ClueBot NG. (41331) (Bot))* (undo)
- (cur | prev) 22:18, 18 November 2010 98.228.34.169 (talk) (3,521 bytes) (undo)
- (cur | prev) 23:04, 14 November 2010 96.51.77.240 (talk) (3,508 bytes) *(→See also)* (undo)
- (cur | prev) 21:03, 29 September 2010 RadioFan (talk | contribs) (3,518 bytes) *(Reverted edit by 98.183.181.231 identified as vandalism using STiki)* (undo)
- (cur | prev) 20:56, 29 September 2010 98.183.181.231 (talk) (3,547 bytes) *(→See also)* (undo)
- (cur | prev) 03:59, 29 September 2010 Nihil novi (talk | contribs) (3,518 bytes) *(see also)* (undo)
- (cur | prev) 08:38, 28 September 2010 Shadowjams (talk | contribs) **m** (3,470 bytes) *(Reverted edits by 124.107.158.76 (talk) to last revision by Pharaoh of the Wizards (HG))* (undo)
- (cur | prev) 08:38, 28 September 2010 124.107.158.76 (talk) (3,480 bytes) *(→References)* (undo)

Wikipedia's Revision History

Paraphrase

From Wikipedia, the free encyclopedia
(Difference between revisions)

Revision as of 02:55, 15 October 2007 (edit)	Revision as of 18:04, 15 October 2007 (edit) (undo)
99.244.101.170 (talk)	88.74.162.233 (talk)
(→Notes)	(reverting vandalism)
← Previous edit	Next edit →

Line 4:	Line 4:
== Characteristics of a well-done paraphrase ==	== Characteristics of a well-done paraphrase ==
- *It is a summary.	+ *It is **not** a summary.
- *It does contain most of the words or phrases from the original.	+ *It does **not** contain most of the words or phrases from the original **([[plagiarism]])**.
*It includes all minor details from original.	*It includes all minor details from original.
*The meaning of the writing being paraphrased is clearer to the reader than in the original text.	*The meaning of the writing being paraphrased is clearer to the reader than in the original text.
- *It **usually does not restate** the thesis	+ *It **restates** the thesis
- *It is longer than the original **quote**.	+ *It is **usually** longer than the original.
==Example==	==Example==

Wikipedia's Revision History

New features Log in / create accou

WIKIPEDIA
The Free Encyclopedia

Article Discussion Read Edit View history Search

Maltase

From Wikipedia, the free encyclopedia
(Difference between revisions)

Main page
Contents
Featured content
Current events
Random article
Interaction
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Donate to Wikipedia
Help
Toolbox
Languages
Deutsch

Revision as of 04:46, 25 January 2007 (edit)
Arcadian (talk | contribs)
({{Glycoside hydrolases}} and mesh link)
← Previous edit

Revision as of 01:39, 15 January 2008 (edit) (undo)
Drphilharmonic (talk | contribs)
(logic, grammar, syntax)
Next edit →

(12 intermediate revisions not shown)

Line 1:

\- The maltase works **like any** other **enzyme**, with the [[substrate (biochemistry)|substrate]] (maltose) binding with the [[active site]]. **When** the maltose **had bound** with the maltase, the former is hydrolysed, that is **to say** it is split into its component parts, i.e. two molecules of α-glucose. This is done by breaking the [[glycosidic bond]] between the 'first' carbon of one glucose, and the 'fourth' carbon of the other (a 1-4 bond).

Line 1:

\+ The maltase works **in the same manner as** other **enzymes**, with the [[substrate (biochemistry)|substrate]] (maltose) binding with the [[active site]]. **After** the maltose **binds** with the maltase, the former is hydrolysed, that is**,** it is split into its component parts, i.e.**,** two molecules of α-glucose. This is done by breaking the [[glycosidic bond]] between the 'first' carbon of one glucose, and the 'fourth' carbon of the other (a 1-4 bond).

Wikipedia's Revision History

Resource: French *Wikipedia*

Large quantity of data for research, focus on Wikipedia's revision history

Hypothesis

Modifications which preserve meaning can be considered as paraphrases

Goal: automatic classification

Modifications which preserve meaning

vs

Modifications which alter meaning

Uses of Wikipedia's revision history in NLP

- Extraction of lexical simplifications [Yatskar et al., 2010]
- Spelling correction
  [Nelken and Yamangil, 2008, Max and Wisniewski, 2010]
- Textual entailment [Zanzotto and Pennacchiotti, 2010]

Plan

The WiCoPaCo Corpus

- Local modifications in context mined from **Wikipedia's revision history**
- Corpus freely available from:
  http://wicopaco.limsi.fr
  (**French** data: 408,816 modifications)
- XML encoded with links to Wikipedia info (article, revision, contributor, comments, etc.)
- Allows for external annotation of the data

Wikipedia Correction and Paraphrase Corpus

XML Structure

```
<modifs>
<modif id= "316021" wp_page_id= "307637" wp_before_rev_id=
"12730172" wp_after_rev_id= "12811994" wp_user_id= "0"
wp_user_num_modif= "1096911" wp_comment= "Dernière version :
Flight Simulator X">
<before>Le jeu <m num_words= "1">ne</m> est compatible
qu'avec Windows XP ou Windows Vista.</before>
<after>Le jeu <m num_words= "1">n'</m> est compatible qu'avec
Windows XP ou Windows Vista.</after>
</modif>
</modifs>
```

A Typology of Local Modifications

- Objectives
  - Cover all observable phenomena in the WiCoPaCo corpus
- Fundamental Principle
  - Semantic dichotomy: weak semantic differences vs. strong semantic differences

Weak Semantic Differences: Surface Corrections

- Typographical corrections

  Le triceps brachial est un muscle extenseur de l'avant bras → avant-bras sur le bras.

  eng: The triceps is an extensor muscle of the [forearm] on the arm.
- Non-word spelling corrections

  Ces trois parties se rejoingnent → rejoignent pour former une épaisse masse.

  eng: These three parts [come together] to form a thick mass.
- Context-dependent word corrections (real words)

  L'anathème pour le pêcheur → pécheur : ce dernier est privé de sépulture chrétienne.

  eng: A curse for the [fisherman → sinner]: he is deprived of Christian burial.

Weak Semantic Differences: Rephrasings

- Lexical rephrasings
  L'implémentation → La mise en œuvre de l'algorithme...
  eng: The [implementation] of the algorithm...
- Syntactical rephrasings
  Un infomercial pseudo-scientifique en exposant → qui expose grossièrement...
  eng: A pseudoscientific infomercial [in roughly outlining → which roughly outlines]...
- Semantic rephrasings
  Il fonde le [journal → quotidien] francophone "Le Tunisien" en 1907.
  eng: He founded the French-speaking [newspaper → daily paper] "Le Tunisien" in 1907.

Strong Semantic Differences: Factual Corrections

- Antonymy

  Un catalyseur solide (phase liquide → solide) avec de l'hydrogène (phase gazeuse).

  eng: A solid catalyst ([liquid → solid] phase) with hydrogen (gas phase).

- No apparent semantic link

  Représente pour eux l'Occident chrétien → la supériorité de la race celto-germanique.

  eng: represents for them [the Christian West → the superiority of the Celtic-Germanic race].

Strong Semantic Differences: Vandalism

- Obvious vandalism
  L'Autriche a été occupée par → psh ! ! ar les Romains.
  eng: Austria was occupied [by → bsh ! ! y] the Romans.

- Subtle vandalism
  Devant la Cour de Cassation → Castration...
  eng: In front of the Court of [Cassation → Castration]...

Plan

Design of the annotation schema

- Objectives
  - Distinguish rephrasings (paraphrases) from surface corrections and strong semantic differences
  - Assess the difficulty of manually identifying paraphrases within local modifications.

- Annotation principles
  - An annotation covers the entire segment identified as a local modification
  - Several labels can be assigned to the same modification segment

Annotation schema

- Surface corrections
  Modifications which aim at making the text compliant with language standards
- Rephrasings
  Different kinds of paraphrases: reformulations, precisions and simplifications
- Strong semantic variations
  Vandalism and factual corrections
- Misalignments
  The local modifications identified present a default in their alignment.

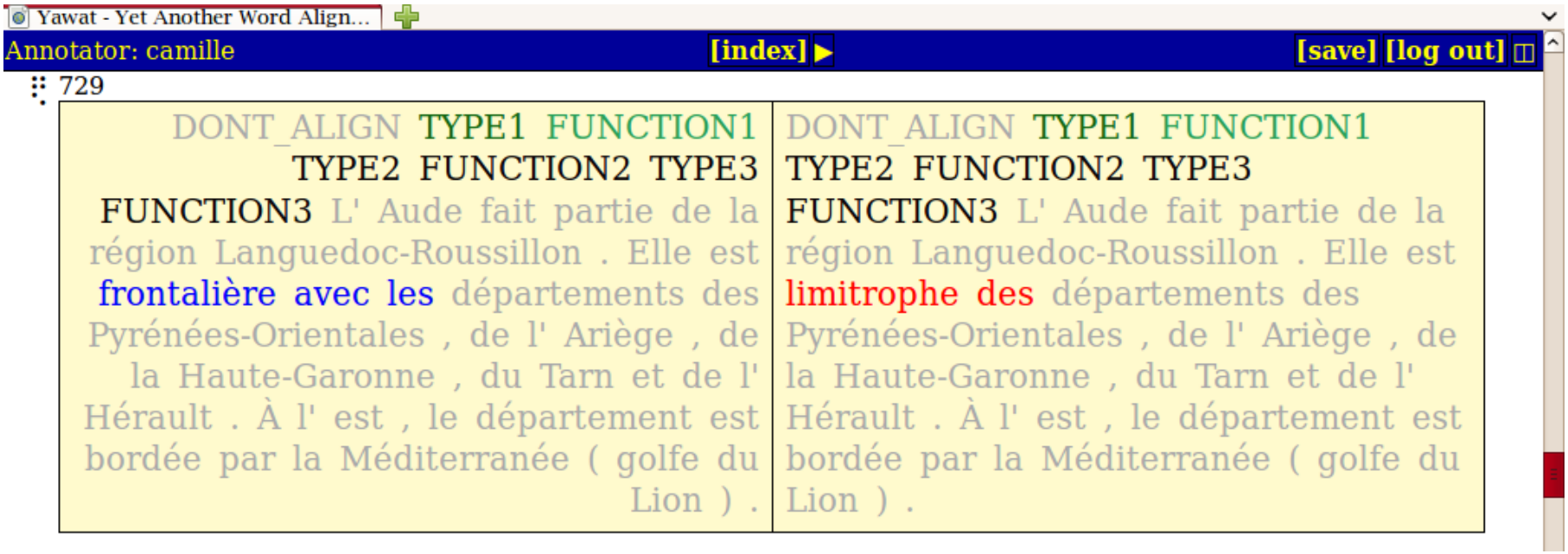
YAWAT: Yet Another Word Alignment Tool

Yet Another Word Alignment Tool

- Characteristics
  - Written by Ulrich Germann [Germann, 2008]
  - Targeted at the alignment of *bilingual* parallel texts
  - Dynamic Web application
  - JavaScript
- Adaptation
  - Use of our own annotation schema
  - No re-alignment of the segments

Yet Another Word Alignment Tool

Web Inteface



Annotation of a sub-part of WiCoPaCo corpus

- Method
  - 200 pairs of modification segments
  - Filtered version the WiCoPaCo corpus: only modification segments with a Levenshtein edit distance of at least 4 were considered for annotation
  - 4 annotators
- Kappa κ
  - Strong semantic variation: substantial agreement
  - Other classes: moderate agreement
- Quantification of the phenomena
  - Rephrasings have the largest number of occurrences (132), followed by strong semantic variations (107)
  - Only few misalignments (20)

Difficulties of the manual annotation

- Several phenomena may occur simultaneously
- The sentential context provided by the WiCoPaCo corpus is sometimes not sufficient to make a decision about a specific modification type
- Correctly typing a modification may necessitate some external knowledge about the contributor's intentions

Plan

Preliminary study

- Resource : development corpus
  - Taken from the *Multitrad* dataset
    Built by collecting several translations for the same input text during a web-based experiment [Bouamor, 2010].
- Method : Transformation rules
  - Based on the *TreeTagger*

Fast Term Recognizer

- Description
  - Developed by C. Jacquemin at LIMSI [Christian Jacquemin, 1994]
  - Multilingual term indexing
  - Term variant recognition
- Associated resources
  - Morphological families for French
  - Semantic links for French

Design of paraphrasing metarules

Metarule NAtoVA(X1 → N1 A1)= X1 → V1 {ART?|PRON?|PREP?} N A1:
<N1 root> = <V1 root>
<X1 metaLabel> = 'XX'.

analyse statistique	→	*analyser des données statistiques*
statistical analysis	→	analyse statistical data
N1root A1	→	V1root ART N A1
consommation régulière	→	*consommer de façon régulière*
regular consuming	→	consume regularly
N1root A1	→	V1root PREP N A1

Results

- **83 metarules have been developed**
- Coverage
  - `Fastr` was able to identify 185 paraphrase candidates in a sub-part of the Multitrad corpus (206 sentence pairs) which was not used for rule development
- WiCoPaCo dataset
  - Corpus of positive and negative paraphase examples (200 of each type) from the WiCoPaCo corpus
  - 31 pairs of candidate paraphrases are identified in the positive corpus
    - 22 (70%) cover the whole modification segment
    - 7 (22.5%) correspond to a subpart of the modification
  - 4 pairs of candidate paraphrases in the negative corpus

Limitations

- Necessitates a large number of metarules
- Rigid formalism
- Dependent on the coverage and quality of the associated morphological and semantic resources

Plan

Conclusion and perspectives

Conclusion

- Detailed typology of the local modification phenomena which are present in Wikipedia's revision history
- Manual annotation study
- Evaluation of a rule-based approach to paraphrase identification

Perspectives

- Combine rule-based and statistical paraphrase identification techniques
- Constitute a large-scale resource of paraphrases automatically extracted from Wikipedia's revision history

Thanks for your attention.

Thanks for listening.

I thank you for listening to me.

I thank you for your attention.

References I


Bouamor, H. (2010).
Construction d'un corpus de paraphrases d'énoncés par traduction multiple multilingue.
In *Actes de RÉCITAL 2010*, Montréal, Canada.


Christian Jacquemin (1994).
Recycling terms into a partial parser.
In *Proceedings of the fourth conference on Applied natural language processing*, pages 113–118, Stuttgart, Germany.


Germann, U. (2008).
Yawat: Yet Another Word Alignment Tool.
In *Proceedings of the ACL-08: HLT Demo Session (Companion Volume)*, pages 20–23.


Max, A. and Wisniewski, G. (2010).
Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History.
In *Proceedings of LREC 2010*, Valletta, Malta.

References II

Nelken, R. and Yamangil, E. (2008).
Mining Wikipedia's Article Revision History for Training Computational Linguistic Algorithms.
In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence:An Evolving Synergy*, pages 31–36.

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010).
For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia.
In *Proceedings of the NAACL*, pages 365–368.

Zanzotto, F. M. and Pennacchiotti, M. (2010).
Expanding textual entailment corpora from Wikipedia using co-training.
In *Proceedings of the 2nd Workshop on Collaboratively Constructed Semantic Resources*.