Chunking Shared Task

Text chunking consists of dividing a text in syntactically correlated parts of words. For example, the sentence “He reckons the current account deficit will narrow to only # 1.8 billion in September” can be divided as follows:

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September] .

Text chunking is an intermediate step towards full parsing. It was the shared task for CoNLL-2000. Training and test data for this task is available. This data consists of the same partitions of the Wall Street Journal corpus (WSJ) as the widely used data for noun phrase chunking: sections 15-18 as training data (211727 tokens) and section 20 as test data (47377 tokens). The annotation of the data has been derived from the WSJ corpus by a program written by Sabine Buchholz from Tilburg University, The Netherlands.

The goal of this task is to come forward with machine learning methods which after a training phase can recognize the chunk segmentation of the test data as well as possible. The training data can be used for training the text chunker. The chunkers will be evaluated with the F rate, which is a combination of the precision and recall rates: F = 2*precision*recall / (recall+precision) [Rij79]. The precision and recall numbers will be computed over all types of chunks.

In this task eleven machine learning-based methods were proposed, along with a baseline method which selects the chunk tag which is most frequently associated with the current part-of-speech tag. Among the non-baseline methods is a Maximum Entropy model for chunking developed by Robert Koeling [Koe00].

Your task will be to implement (i) the baseline method and (ii) the MaxEnt model [Koe00] and establish parity with the reported F-scores. Of course, in case you want to implement any other non-baseline method, you are welcome to do so.

The information about the shared task, including the proposed approaches, can be found at the following link: https://www.clips.uantwerpen.be/conll2000/chunking/. For the sake of completeness the references are reproduced here. [Koe00] is the fourth reference.

References

This reference section contains two parts: first the papers from the shared task session at CoNLL-2000 and then the other related publications.

CoNLL-2000 Shared Task Papers

- **[TB00]**
  Erik F. Tjong Kim Sang and Sabine Buchholz, Introduction to the CoNLL-2000 Shared Task: Chunking. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.
  [abstract] [ps] [pdf]
- **[Dej00]**
  Hervé Déjean, Learning Syntactic Structures with XML. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.
  [ps] [pdf] [test data output]
- **[Joh00]**
  Christer Johansson, A Context Sensitive Maximum Likelihood Approach to Chunking. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.
  [ps] [pdf] [test data output]
- **[Koe00]**
  Rob Koeling, Chunking with Maximum Entropy Models. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.
  [ps] [pdf] [test data output]
- **[KM00]**
  Taku Kudoh and Yuji Matsumoto, Use of Support Vector Learning for Chunk Identification. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.
  [ps] [pdf] [test data output]
- **[Osb00]**
  Miles Osborne, Shallow Parsing as Part-of-Speech Tagging. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.
  [abstract] [ps] [pdf] [test data output]
- **[PMP00]**
  Ferran Pla, Antonio Molina and Natividad Prieto, Improving Chunking by Means of Lexical-Contextual Information in Statistical Language Models. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.
  [ps] [pdf] [test data output]
- **[TKS00]**
  Erik F. Tjong Kim Sang, Text Chunking by System Combination. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.
  [ps] [pdf] [test data output]
- **[Hal00]**
  Hans van Halteren, Chunking with WPDV Models. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.
  [ps] [pdf] [test data output]
- **[VB00]**
  Jorn Veenstra and Antal van den Bosch, Single-Classifier Memory-Based Phrase Chunking. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.
  [ps] [pdf] [test data output]

- **[VD00]**
  Marc Vilain and David Day, Phrase Parsing with Rule Sequence Processors: an Application to the Shared CoNLL Task. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.
  [ps] [pdf] [test data output]
- **[ZST00]**
  GuoDong Zhou, Jian Su and TongGuan Tey, Hybrid Text Chunking. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.
  [abstract] [ps] [pdf] [test data output]

Other related publications

- **[Abn91]**
  Steven Abney, Parsing By Chunks. In: Robert Berwick and Steven Abney and Carol Tenny, "Principle-Based Parsing", Kluwer Academic Publishers, 1991.
  http://whorf.sfs.nphil.uni-tuebingen.de/~abney/Abney_90e.ps.gz
- **[Bel01]**
  Anja Belz, Optimisation of corpus-derived probabilistic grammars, In: "Corpus Linguistics 2001", Lancaster, UK, 2001.
  http://www.clips.uantwerpen.be/lcg/ps/belz.cl2001.ps.gz
- **[BVD99]**
  Sabine Buchholz, Jorn Veenstra and Walter Daelemans, Cascaded Grammatical Relation Assignment. In: "Proceedings of EMNLP/VLC-99", University of Maryland, USA, 1999.
  ftp://ilk.kub.nl/pub/papers/ilk.9908.ps.gz
- **[CM03]**
  Xavier Carreras and Lluís Màrquez, Phrase Recognition by Filtering and Ranking with Perceptrons. In "Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP-2003", Borovets, Bulgaria, 2003.
  http://www.lsi.upc.es/~nlp/papers/2003/ranlp2003-cm.ps.gz
- **[Dej02]**
  Hervé Déjean, Learning Rules and Their Exceptions. In Journal of Machine Learning Research, volume 2 (March), 2002, pp. 669-693.
  http://www.ai.mit.edu/projects/jmlr/papers/volume2/dejean02a/dejean02a.pdf
- **[FHN00]**
  Radu Florian, John C. Henderson and Grace Ngai, Coaxing Confidences from an Old Friend: Probabilistic Classifications from Transformation Rule Lists. In: "Proceedings of EMNLP 2000", Hong Kong, 2000.
  http://arXiv.org/ps/cs/0104020
- **[KM01]**
  Taku Kudoh and Yuji Matsumoto, Chunking with Support Vector Machines, In: "Proceedings of NAACL 2001", Pittsburgh, PA, USA, 2001.
  http://cactus.aist-nara.ac.jp/~taku-ku/publication/naacl2001.ps
- **[Meg02]**
  Beáta Megyesi, Shallow Parsing with PoS Taggers and Linguistic Features. In

Journal of Machine Learning Research, volume 2 (March), 2002, pp. 639-668.
http://www.ai.mit.edu/projects/jmlr/papers/volume2/megyesi02a/megyesi02a.pdf

- **[MP02]**
  Antonio Molina and Ferran Pla, Shallow Parsing using Specialized HMMs, In Journal of Machine Learning Research, volume 2 (March), 2002, pp. 595-613.
  http://www.ai.mit.edu/projects/jmlr/papers/volume2/molina02a/molina02a.pdf
- **[NF01]**
  Grace Ngai and Radu Florian. Transformation Based Learning in the Fast Lane. In: "Proceedings of NAACL 2001", Pittsburgh, PA, USA, 2001.
  http://nlp.cs.jhu.edu/~rflorian/papers/naacl01.ps
- **[Osb02]**
  Miles Osborne, Shallow Parsing using Noisy and Non-Stationary Training Material. In Journal of Machine Learning Research, volume 2 (March), 2002, pp. 695-719.
  http://www.ai.mit.edu/projects/jmlr/papers/volume2/osborne02a/osborne02a.pdf
- **[RM95]**
  Lance A. Ramshaw and Mitchell P. Marcus, Text Chunking Using Transformation-Based Learning. In: "Proceedings of the Third ACL Workshop on Very Large Corpora", Cambridge MA, USA, 1995.
  ftp://ftp.cis.upenn.edu/pub/chunker/wvlcbook.ps.gz
- **[Rat98]**
  Adwait Ratnaparkhi, "Maximum Entropy Models for Natural Language Ambiguity Resolution". PhD thesis, University of Pennsylvania, 1998.
  ftp://ftp.cis.upenn.edu/pub/ircs/tr/98-15/98-15.ps.gz
- **[Rij79]**
  C.J. van Rijsbergen, "Information Retrieval". Buttersworth, 1979.
- **[SP03]**
  Fei Sha and Fernando Pereira. Shallow Parsing with Conditional Random Fields. In: Proceedings of HLT-NAACL 2003, Edmonton, Canada, 2003, pp. 213-220.
  http://www.cis.upenn.edu/~feisha/pubs/shallow03.pdf
- **[TKS02]**
  Erik F. Tjong Kim Sang, Memory-Based Shallow Parsing, In Journal of Machine Learning Research, volume 2 (March), 2002, pp. 559-594.
  http://arXiv.org/abs/cs.CL/0204049
- **[Vee99]**
  Jorn Veenstra. Memory-Based Text Chunking, In: Nikos Fakotakis (ed), "Machine learning in human language technology", workshop at ACAI 99, Chania, Greece, 1999.
  http://ilk.kub.nl/~ilk/papers/ACAI.ps
- **[ZDJ01]**
  Tong Zhang, Fred Damerau and David Johnson, Text Chunking using Regularized Winnow. In: Proceedings of ACL-2001, Toulouse, France, 2001.
- **[ZDJ02]**
  Tong Zhang, Fred Damerau and David Johnson, Text Chunking based on a Generalization of Winnow. In Journal of Machine Learning Research, volume 2

(March), 2002, pp. 615-637.
http://www.ai.mit.edu/projects/jmlr/papers/volume2/zhang02c/zhang02c.pdf