WIKIPEDIA

Named-entity recognition

Named-entity recognition (**NER**) (also known as **entity identification**, **entity chunking** and **entity extraction**) is a subtask of information extraction that seeks to locate and classify named entity mentions in unstructured text into pre-defined categories such as the person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

Most research on NER systems has been structured as taking an unannotated block of text, such as this one:

> Jim bought 300 shares of Acme Corp. in 2006.

And producing an annotated block of text that highlights the names of entities:

> $[\text{Jim}]_{\text{Person}}$ bought 300 shares of $[\text{Acme Corp.}]_{\text{Organization}}$ in $[\text{2006}]_{\text{Time}}$.

In this example, a person name consisting of one token, a two-token company name and a temporal expression have been detected and classified.

State-of-the-art NER systems for English produce near-human performance. For example, the best system entering MUC-7 scored 93.39% of F-measure while human annotators scored 97.60% and 96.95%.[1][2]

Contents

Named-entity recognition platforms

Notable NER platforms include:

- GATE supports NER across many languages and domains out of the box, usable via a graphical interface and a Java API.
- OpenNLP includes rule-based and statistical named-entity recognition.
- SpaCy features fast statistical NER as well as an open-source named-entity visualizer.

Problem definition

In the expression *named entity*, the word *named* restricts the task to those entities for which one or many strings, such as words or phrases, stands (fairly) consistently for some referent. This is closely related to rigid designators, as defined by Kripke[3][4], although in practice NER deals with many names and referents that are not philosophically "rigid". For instance, the *automotive company created by Henry Ford in 1903* can be referred to as *Ford* or *Ford Motor Company*, although "Ford" can refer to many other entities as well (see Ford). Rigid designators include proper names as well as terms for certain biological species and substances,[5] but exclude pronouns (such as "it"; see coreference resolution), descriptions that pick out a referent by its properties (see also De dicto and de re), and names for kinds of things as opposed to individuals (for example "Bank").

Full named-entity recognition is often broken down, conceptually and possibly also in implementations,[6] as two distinct problems: detection of names, and classification of the names by the type of entity they refer to (e.g. person, organization, location and other[7]). The first phase is typically simplified to a segmentation problem: names are defined to be contiguous spans of tokens, with no nesting, so that "Bank of America" is a single name, disregarding the fact that inside this name, the substring "America" is itself a name. This segmentation problem is formally similar to chunking. The second phase requires choosing an ontology by which to organize categories of things.

Temporal expressions and some numerical expressions (i.e., money, percentages, etc.) may also be considered as named entities in the context of the NER task. While some instances of these types are good examples of rigid designators (e.g., the year 2001) there are also many invalid ones (e.g., I take my vacations in “June”). In the first case, the year *2001* refers to the *2001st year of the Gregorian calendar*. In the second case, the month *June* may refer to the month of an undefined year (*past June*, *next June*, *every June*, etc.). It is arguable that the named entity definition is loosened in such cases for practical reasons. The definition of the term *named entity* is therefore not strict and often has to be explained in the context in which it is used.[8]

Certain hierarchies of named entity types have been proposed in the literature. BBN categories, proposed in 2002, is used for Question Answering and consists of 29 types and 64 subtypes.[9] Sekine's extended hierarchy, proposed in 2002, is made of 200 subtypes.[10] More recently, in 2011 Ritter used a hierarchy based on common Freebase entity types in ground-breaking experiments on NER over social media text.[11]

Formal evaluation

To evaluate the quality of a NER system's output, several measures have been defined. While accuracy on the token level is one possibility, it suffers from two problems: the vast majority of tokens in real-world text are not part of entity names as usually defined, so the baseline accuracy (always predict "not an entity") is extravagantly high, typically >90%; and mispredicting the full span of an entity name is not properly penalized (finding only a person's first name when their last name follows is scored as ½ accuracy).

In academic conferences such as CoNLL, a variant of the F1 score has been defined as follows:[7]

- Precision is the number of predicted entity name spans that line up *exactly* with spans in the gold standard evaluation data. I.e. when [$_{\text{Person}}$ Hans] [$_{\text{Person}}$ Blick] is predicted but [$_{\text{Person}}$ Hans Blick] was required, precision for the predicted name is zero. Precision is then averaged over all predicted entity names.
- Recall is similarly the number of names in the gold standard that appear at exactly the same location in the predictions.
- F1 score is the harmonic mean of these two.

It follows from the above definition that any prediction that misses a single token, includes a spurious token, or has the wrong class, is a hard error and does not contribute to either precision or recall.

Evaluation models based on a token-by-token matching have been proposed.[12] Such models are able to handle also partially overlapping matches, yet fully rewarding only exact matches. They allow a finer grained evaluation and comparison of extraction systems, taking into account also the degree of mismatch in non-exact predictions.

Approaches

NER systems have been created that use linguistic grammar-based techniques as well as statistical models such as machine learning. Hand-crafted grammar-based systems typically obtain better precision, but at the cost of lower recall and months of work by experienced computational linguists [13]. Statistical NER systems typically require a large amount of manually annotated training data. Semisupervised approaches have been suggested to avoid part of the annotation effort.[14][15]

Many different classifier types have been used to perform machine-learned NER, with conditional random fields being a typical choice.[16]

Problem domains

Research indicates that even state-of-the-art NER systems are brittle, meaning that NER systems developed for one domain do not typically perform well on other domains.[17] Considerable effort is involved in tuning NER systems to perform well in a new domain; this is true for both rule-based and trainable statistical systems.

Early work in NER systems in the 1990s was aimed primarily at extraction from journalistic articles. Attention then turned to processing of military dispatches and reports. Later stages of the automatic content extraction (ACE) evaluation also included several types of informal text styles, such as weblogs and text transcripts from conversational telephone speech conversations. Since about 1998, there has been a great deal of interest in entity identification in the molecular biology, bioinformatics, and medical natural language processing communities. The most common entity of interest in that domain has been names of genes and gene products. There has been also considerable interest in the recognition of chemical entities and drugs in the context of the CHEMDNER competition, with 27 teams participating in this task.[18]

Current challenges and research

Despite the high F1 numbers reported on the MUC-7 dataset, the problem of Named Entity Recognition is far from being solved. The main efforts are directed to reducing the annotation labor by employing semi-supervised learning,[14][19] robust performance across domains[20][21] and scaling up to fine-grained entity types.[10][22] In recent years, many projects have turned to crowdsourcing, which is a promising solution to obtain high-quality aggregate human judgments for supervised and semi-supervised machine learning approaches to NER.[23] Another challenging task is devising models to deal with linguistically complex contexts such as Twitter and search queries.[24]

There are some researchers who did some comparisons about the NER performances from different statistical models such as HMM (Hidden Markov Model), ME (Maximum Entropy), and CRF (Conditional Random Fields) and feature sets.[25] And some researchers recently proposed graph-based semi-supervised learning model for language specific NER tasks.[26]

A recently emerging task of identifying "important expressions" in text and cross-linking them to Wikipedia[27][28][29] can be seen as an instance of extremely fine-grained named entity recognition, where the types are the actual Wikipedia pages describing the (potentially ambiguous) concepts. Below is an example output of a Wikification system:

```
<ENTITY url="http://en.wikipedia.org/wiki/Michael_I._Jordan"> Michael Jordan </ENTITY> is a professor at <ENTITY
url="http://en.wikipedia.org/wiki/University_of_California,_Berkeley"> Berkeley </ENTITY>
```

Another field that has seen progress but remains challenging is the application of NER to Twitter and other microblogs.[30]

See also

- Coreference resolution
- Entity linking (aka named entity normalization, entity disambiguation)
- Information extraction
- Knowledge extraction
- Controlled vocabulary
- Onomastics
- Record linkage
- Smart tag (Microsoft)

References

1. Elaine Marsh, Dennis Perzanowski, "MUC-7 Evaluation of IE Technology: Overview of Results", 29 April 1998 PDF (http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/marsh_slides.pdf)
2. MUC-07 Proceedings (Named Entity Tasks) (http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html#named)
3. Kripke, Saul (1971). M.K. Munitz, ed. *Identity and Necessity*. New York: New York University Press. pp. 135–64.
4. LaPorte, Joseph, *Rigid Designators* (https://plato.stanford.edu/entries/rigid-designators/)
5. Nadeau, David; Sekine, Satoshi (2007). *A survey of named entity recognition and classification* (http://nlp.cs.nyu.edu/sekine/papers/li07.pdf) (PDF). Lingvisticae Investigationes.
6. Carreras, Xavier; Màrquez, Lluís; Padró, Lluís (2003). *A simple named entity extractor using AdaBoost*. CoNLL.
7. Tjong Kim Sang, Erik F.; De Meulder, Fien (2003). *Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition* (http://www.aclweb.org/anthology/W03-0419). CoNLL.
8. Named Entity Definition (http://webknox.com/p/named-entity-definition). Webknox.com. Retrieved on 2013-07-21.
9. Brunstein, Ada. "Annotation Guidelines for Answer Types" (https://catalog.ldc.upenn.edu/docs/LDC2005T33/BBN-Types-Subtypes.html). *LDC Catalog*. Linguistic Data Consortium. Retrieved 21 July 2013.
10. Sekine's Extended Named Entity Hierarchy (http://nlp.cs.nyu.edu/ene/). Nlp.cs.nyu.edu. Retrieved on 2013-07-21.
11. Ritter, A.; Clark, S.; Mausam; Etzioni., O. (2011). *Named Entity Recognition in Tweets: An Experimental Study* (https://aclweb.org/anthology/D/D11/D11-1141.pdf) (PDF). Proc. Empirical Methods in Natural Language Processing.
12. Esuli, Andrea; Sebastiani, Fabrizio (2010). *Evaluating Information Extraction* (http://nmis.isti.cnr.it/sebastiani/Publications/CLEF10.pdf) (PDF). Cross-Language Evaluation Forum (CLEF). pp. 100–111.
13. Kapetanios, Epaminondas; Tatar, Doina; Sacarea, Christian (2013-11-14). *Natural Language Processing: Semantic Aspects* (https://books.google.com/books?id=YXv6AQAAQBAJ&pg=PA298&lpg=PA298). CRC Press. p. 298. ISBN 9781466584969.
14. Lin, Dekang; Wu, Xiaoyun (2009). *Phrase clustering for discriminative learning* (http://www.aclweb.org/anthology/P/P09/P09-1116.pdf) (PDF). Annual Meeting of the ACL and IJCNLP. pp. 1030–1038.
15. Nothman, Joel; et al. (2013). "Learning multilingual named entity recognition from Wikipedia". *Artificial Intelligence*. **194**: 151–175. doi:10.1016/j.artint.2012.03.006 (https://doi.org/10.1016%2Fj.artint.2012.03.006).
16. Jenny Rose Finkel; Trond Grenager; Christopher Manning (2005). *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling* (http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf) (PDF). 43rd Annual Meeting of the Association for Computational Linguistics. pp. 363–370.
17. Poibeau, Thierry; Kosseim, Leila (2001). "Proper Name Extraction from Non-Journalistic Texts". *Language and Computers*. **37** (1): 144–157.

18. Krallinger, M; Leitner, F; Rabal, O; Vazquez, M; Oyarzabal, J; Valencia, A. "Overview of the chemical compound and drug name recognition (CHEMDNER) task". *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 2*. pp. 6–37.
19. Turian, J., Ratinov, L., & Bengio, Y. (2010, July). Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 384-394). Association for Computational Linguistics. PDF (http://cogcomp.cs.illinois.edu/papers/TurianRaBe2010.pdf)
20. Ratinov, L., & Roth, D. (2009, June). Design challenges and misconceptions in named entity recognition. (http://cogcomp.cs.illinois.edu/papers/RatinovRo09.pdf) In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (pp. 147-155). Association for Computational Linguistics.
21. Frustratingly Easy Domain Adaptation. (http://www.cs.utah.edu/~hal/docs/daume07easyadapt.pdf)
22. Fine-Grained Named Entity Recognition Using Conditional Random Fields for Question Answering. (http://www.springerlink.com/content/q30m40611u821m2n)
23. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical Natural Language Processing (http://www.jmir.org/2013/4/e73/)
24. Eiselt, Andreas; Figueroa, Alejandro (2013). *A Two-Step Named Entity Recognizer for Open-Domain Search Queries* (https://www.researchgate.net/publication/264083508_A_Two-Step_Named_Entity_Recognizer_for_Open-Domain_Search_Queries). IJCNLP. pp. 829–833.
25. Han, Li-Feng Aaron, Wong, Fai, Chao, Lidia Sam. (2013). Chinese Named Entity Recognition with Conditional Random Fields in the Light of Chinese Characteristics. Proceeding of International Conference of Language Processing and Intelligent Information Systems. M.A. Klopotek et al. (Eds.): IIS 2013, LNCS Vol. 7912, pp. 57–68 [1] (https://link.springer.com/chapter/10.1007%2F978-3-642-38634-3_8#page-1)
26. Han, Li-Feng Aaron, Wong, Zeng, Xiaodong, Derek Fai, Chao, Lidia Sam. (2015). Chinese Named Entity Recognition with Graph-based Semi-supervised Learning Model. In Proceedings of SIGHAN workshop in ACL-IJCNLP. 2015. [2] (http://www.aclweb.org/anthology/W15-3103)
27. Linking Documents to Encyclopedic Knowledge. (http://dl.acm.org/citation.cfm?id=1321475)
28. Learning to link with Wikipedia. (http://www.cs.waikato.ac.nz/~ihw/papers/08-DNM-IHW-LearningToLinkWithWikipedia.pdf)
29. Local and Global Algorithms for Disambiguation to Wikipedia. (http://cogcomp.cs.illinois.edu/papers/RRDA11.pdf)
30. Derczynski, Leon and Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphael Troncy, Johann Petrak, and Kalian Botcheva (2014). “Analysis of named entity recognition and linking for tweets”. *Information Processing and Management* 51(2) : pages 32-49.