# Instructions for assessing availability of code and data in natural language processing papers

1. Examine the abstract. If the questions can be answered on the basis of the abstract, there is no need to examine the full text of the paper. If the questions cannot be answered on the basis of the abstract, then examine the full text of the paper.
2. You are looking for explicit statements.
3. If a web site is given, check to see if it works. If a web site is given but the link is broken, then give the web site address and note that it is broken.
4. If you conclude that code or data is available, then copy and paste the text that led to your conclusion into the spreadsheet. If it is the address of a web site, include the web site address.

## Availability of code

Any of the following should be accepted as evidence that the code is available:

1. The address of a publicly available repository on GitHub or SourceForge. The link is not broken.
2. The address of a web site anywhere else. The link is not broken.
3. A statement that the code is available.

If the paper states that the code is available from the authors, then copy the statement.

## Availability of data

Any of the following should be accepted as evidence that the data is available:

1. The data is from a shared task. Give the name of the shared task.
2. The address of a web site is given. The link is not broken.
3. The data set is one of the very commonly used ones.

If the paper states that the data is available from the authors, then copy the statement.

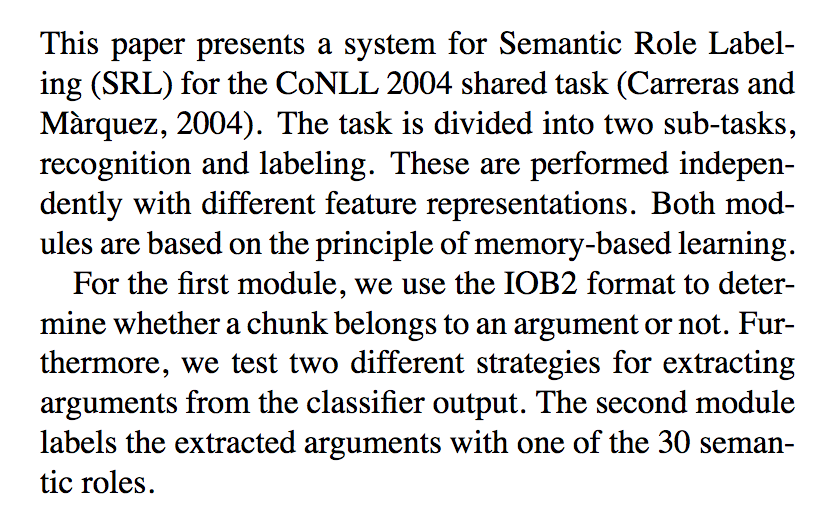
### List of shared tasks

1. BioASQ
2. BioCreative
3. BioNLP-ST
4. CoNLL
5. i2b2
6. LLL (Language in Logic and Learning)
7. MUC (Message Understanding Conference)
8. SEMEVAL
9. SENSEVAL
10. TREC (Text Retrieval Conference)

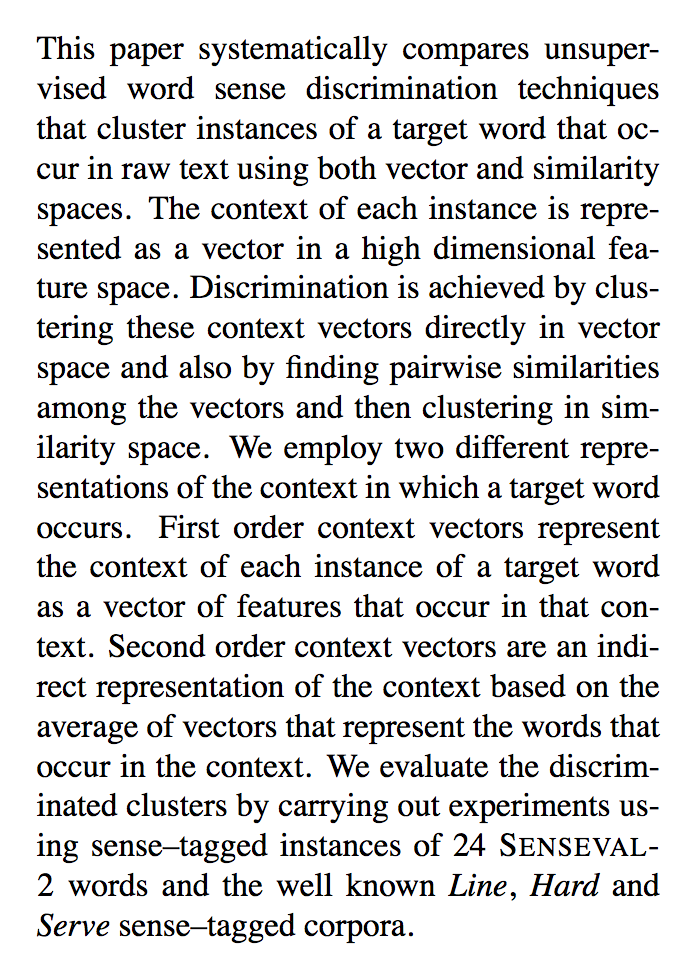
### List of very commonly used data sets

1. GENIA
2. OntoNotes
3. PropBank
4. PTB (Penn Treebank)
5. WSJ (Wall Street Journal)

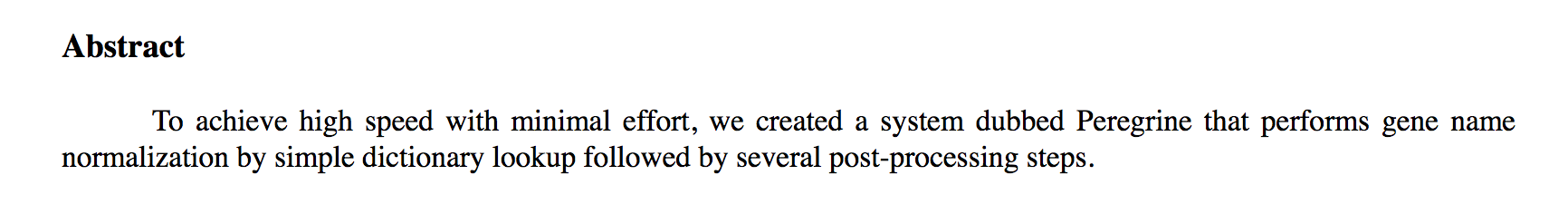
### Examples



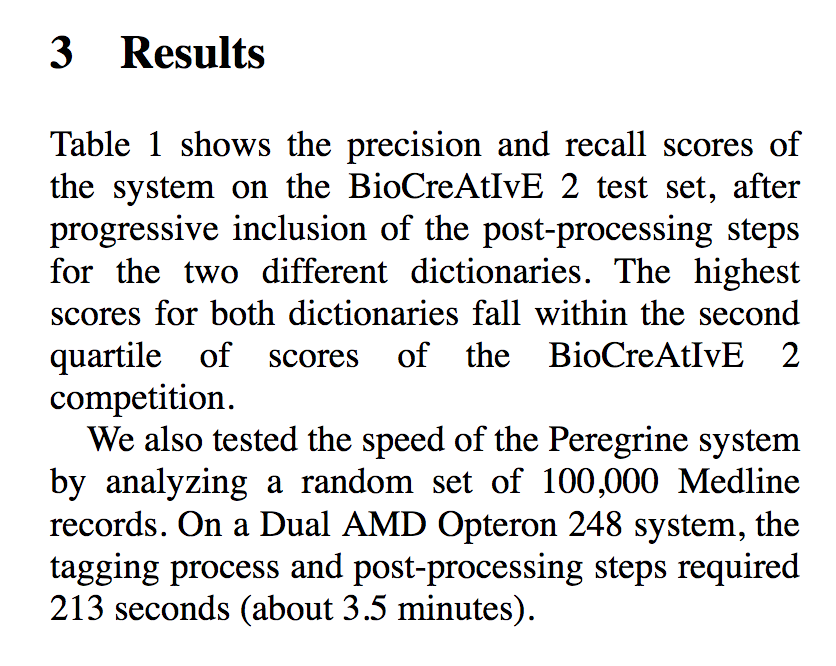
*The paper mentions the CoNLL 2004 shared task, so classify it as having data available.*



*The abstract mentions the SENSEVAL-2 shared task, so classify it as having data available.*



*This (amazingly short) abstract does not mention the data, so the full paper should be checked.*



*This is a tough one. It mentions the BioCreative shared task, and based on that, we know that there is data available. However, it also mentions a set of randomly selected Medline records, and those do not seem to be made available. We need to figure out how to handle this case—for now, make a note of the problem in your spreadsheet.*

## If the paper is behind a paywall

If the paper is behind a paywall, indicate *Paywall* and give the URL for the paper*.*