# 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. **What to do if the full text has more information about the abstract than the**
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.

***What if the tool is available, but not the code****?*

## 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.

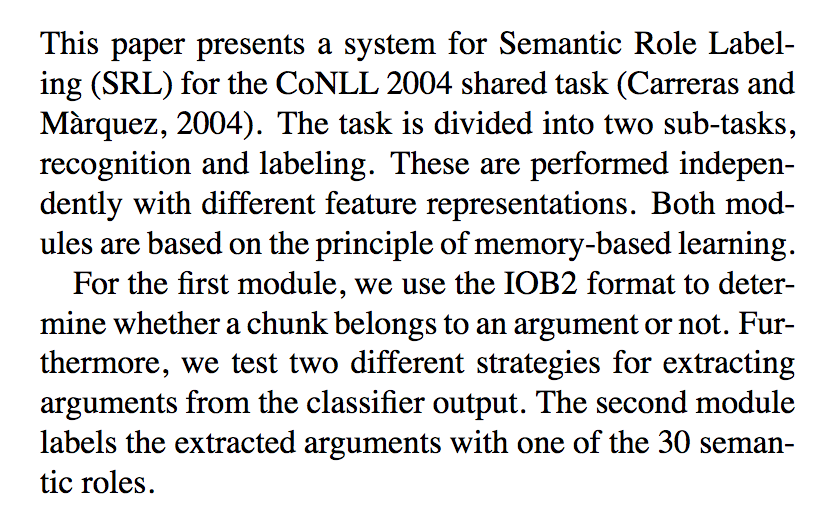
### List of shared tasks

1. BioASQ
2. BioCreative
3. BioNLP-ST
4. CoNLL
5. i2b2
6. LLL (Language in Logic and Learning)
7. MUC (Message Understanding Conference)
8. NIST (National Institute of Standards and Technologies) is associated with many different data sets, mostly TREC (see below), but also others
9. SEMEVAL
10. SENSEVAL
11. TAC (Text Analysis Conference)
12. TREC (Text Retrieval Conference)

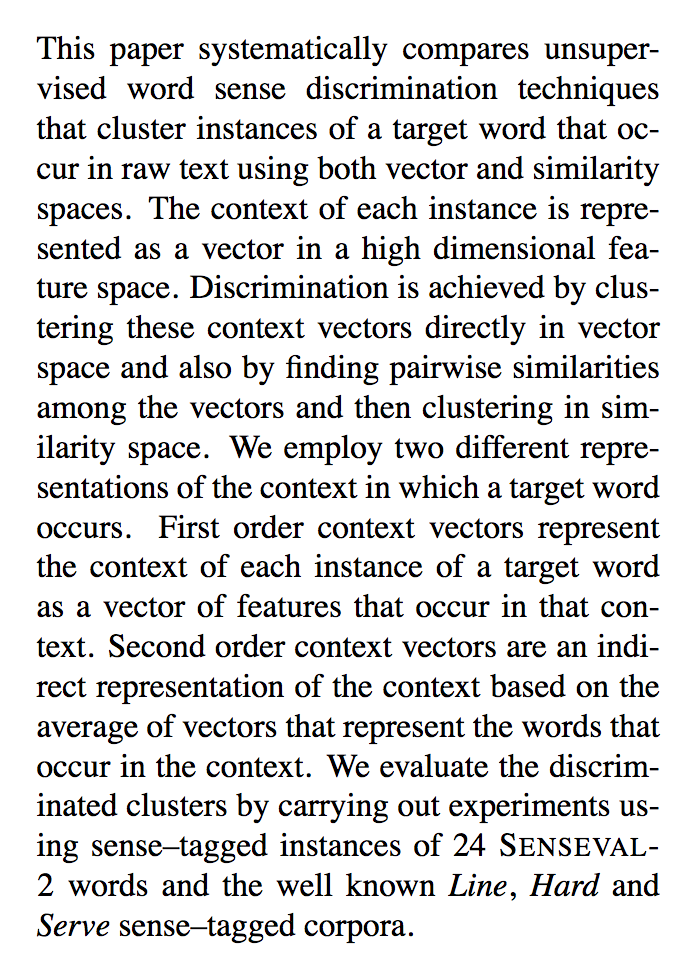
### List of very commonly used data sets

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

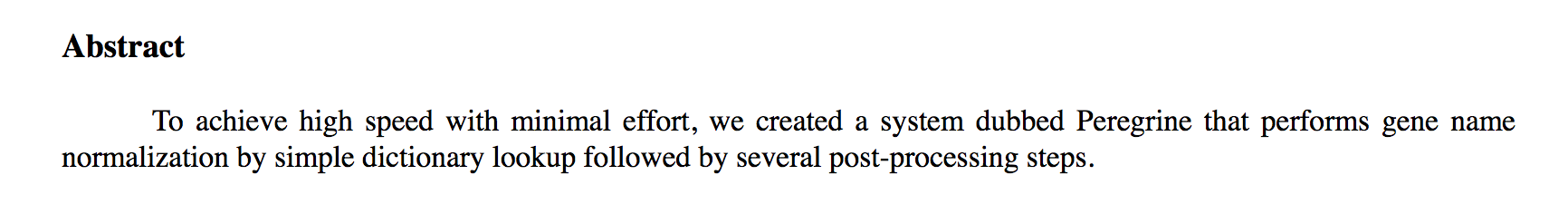
### Examples



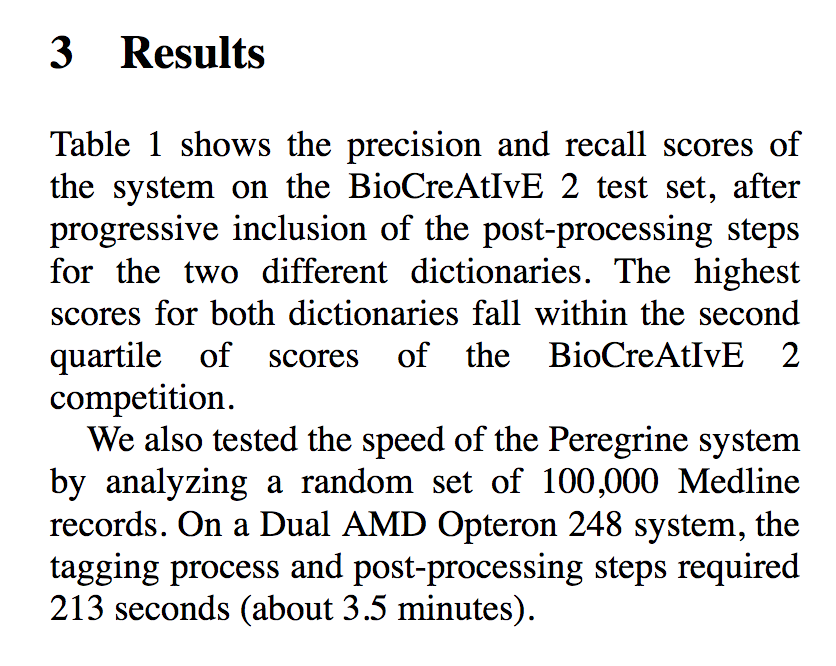
*The paper mentions the CoNLL 2004 shared task, so classify it as having data available.*



*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*.* ***Aurélie: need to expand this to say something about what you tried, and look at the web site that she recommended***

## Kevin

1. Scan the abstract.
2. Search for the string **availab** (catches both **available** and **availability**)
3. Search for the string **download**
4. Search for the strings **data** and **code**
5. Search for the strings **http** and **www**
6. Search for **edu**? **Sourceforge**? **Github**?
7. Scan the Conclusion, then the Methods, then the Results, then the Introduction

Observations: HTML is a lot easier to deal with than PDF, and is often available for PubMed/MEDLINE full-text articles.

# Negacy comments

* duplicates
* what if they mention a data set, but they don’t give a web page?

## Prabha

* Is a model code? (See line 4)
* Is a tag set code? (See line 5)
* If someone says that they used OpenNLP and/or something else that’s available quite clearly, should we say that the code is available? (See line 6)
* Same thing for line 12
* Line 18: what if they say that data is available with IRB approval?
* Line 20: cTAKES?
* Line 23: same thing
* Line 25: what if the algorithm is given in the article? Does that count as the code being available?

## Aurélie

Instructions need examples of copy-paste