

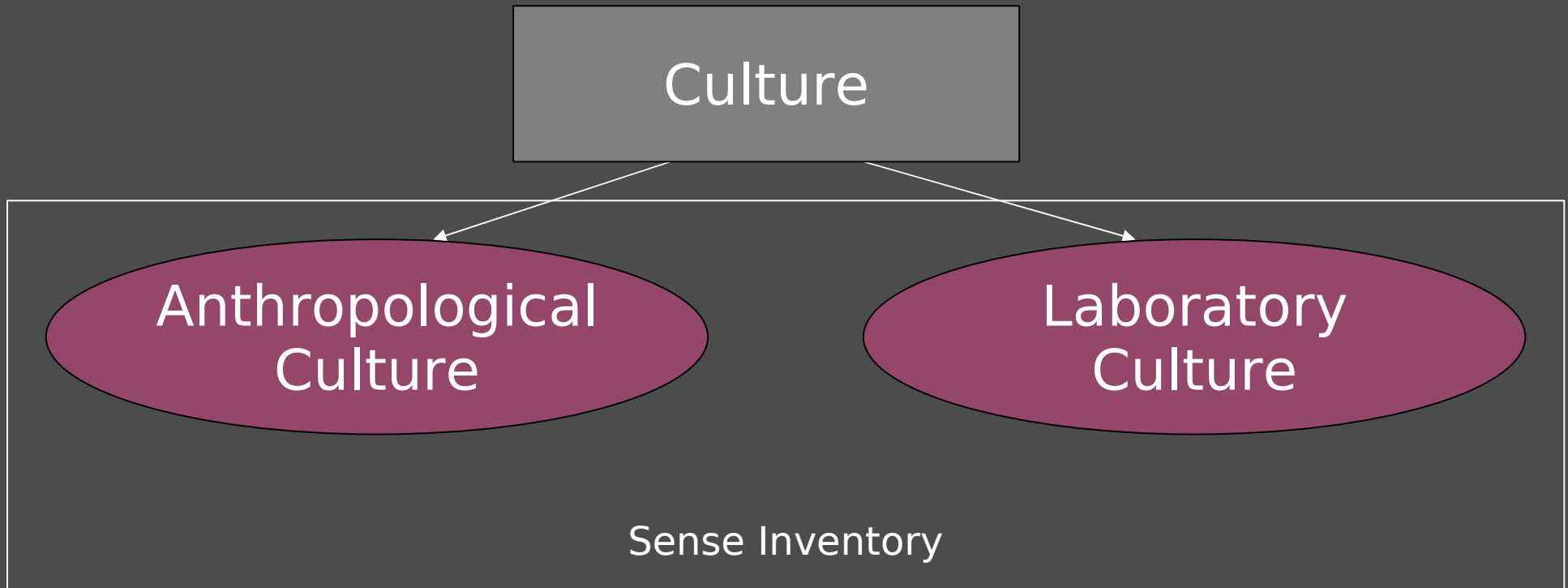
Using UMLS CUIs for WSD in the Biomedical Domain

Bridget T. McInnes¹
Ted Pedersen²
and
John Carlis¹

University of Minnesota Twin Cities¹
and
University of Minnesota Duluth²

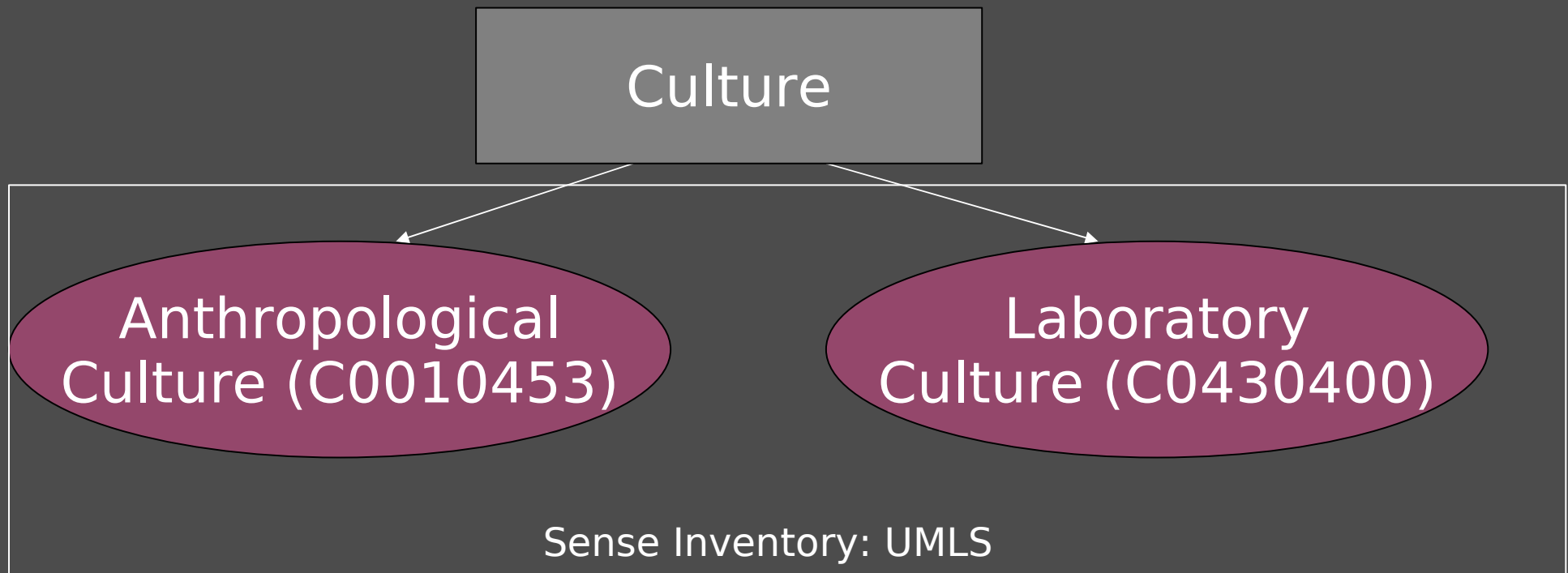
What is WSD?

The *culture* count doubled.



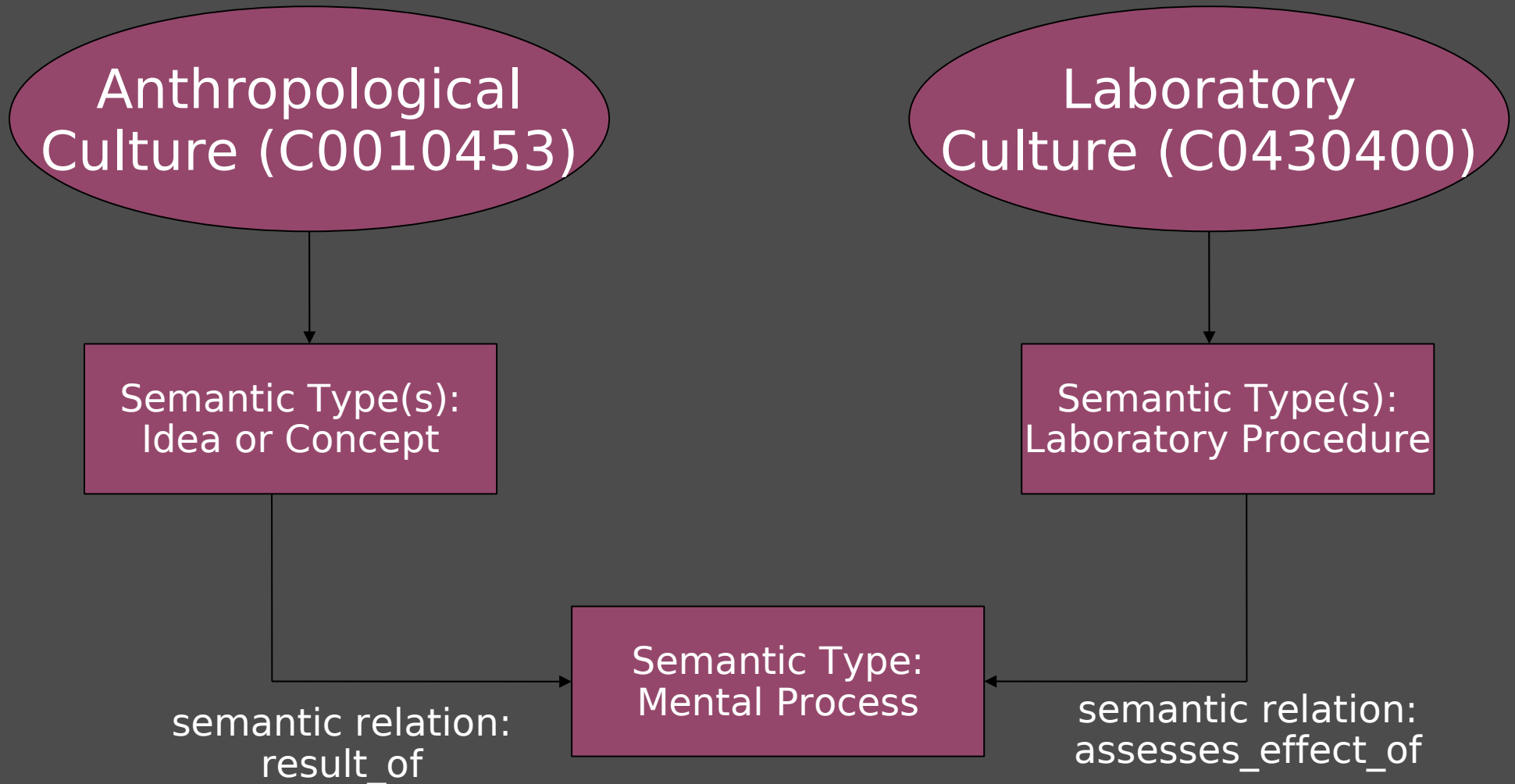
Sense Inventory: UMLS

Unified Medical Language System contains a list of Concept Unique Identifiers (CUIs) which are concepts (senses) associated with a word or term



UMLS: Semantic Network

framework encoded with different semantic and syntactic structures



MetaMap

- Concept mapping system
 - maps text to concepts in the UMLS
 - provides a wealth of information for all words in a document
 - phrasal information
 - Part of speech (POS) of a word
 - CUI of a word
 - Semantic types of a word

Example

- The culture count doubled
 - count
 - CUI: Count (C0750480)
 - semantic type: Idea or Concept (idcn)
 - pos: noun
 - doubled
 - CUI: Duplicate (C0205173)
 - semantic type: Functional Concept (ftcn)
 - pos: verb

Supervised Approaches

- Leroy and Rindflesch 2005
 - Semantic types, semantic relations, part-of-speech, and head information (from MetaMap)
- Joshi, Pedersen and Maclin 2005
 - unigrams
 - in the same sentence as the ambiguous word
 - in the same abstract as the ambiguous word
- Liu, Teller and Friedman 2004
 - unigrams, direction and orientation of unigrams and collocations

Questions

Questions

- Would UMLS CUIs be an improvement over semantic types?

Questions

- Would UMLS CUIs be an improvement over semantic types?
- Would the biomedical specific feature CUIs be an improvement over the more general feature unigrams?

Questions

- Would UMLS CUIs be an improvement over semantic types?
- Would the biomedical specific feature CUIs be an improvement over the more general feature unigrams?
- Would increasing the context window in which surrounding CUIs are found improve the results?

Our supervised approach

- Algorithm:
 - Naïve Bayes from WEKA datamining package using 10 fold cross validation
- Features:
 - UMLS CUIs obtained from MetaMap
 - that occur in the same sentence as the ambiguous word more than one time (s-1-cui)
 - that occur in the same abstract as the ambiguous word more than one time (a-1-cui)

Example

... The *culture* count doubled. The cells multiplied by twice the expected rate ...

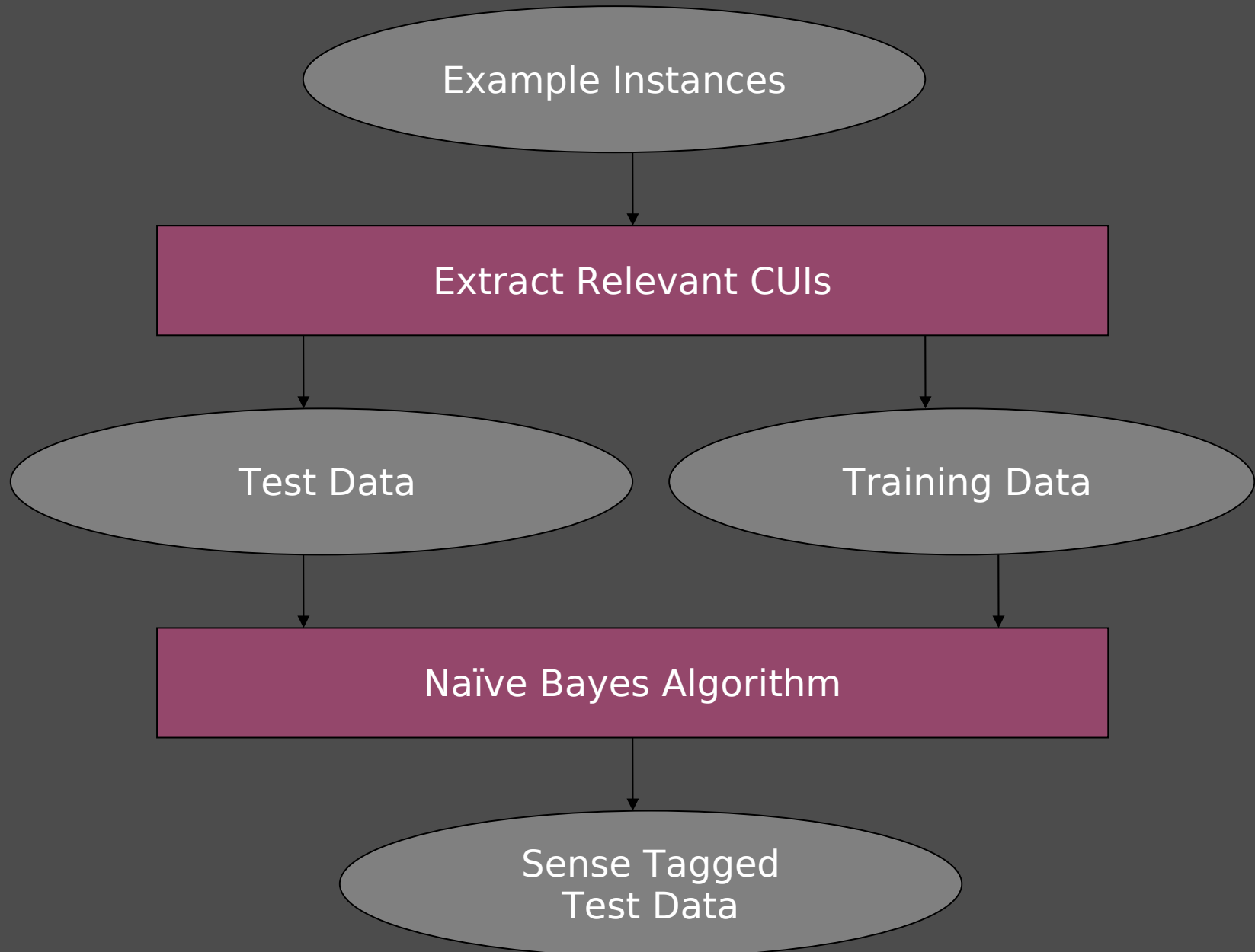
Sentence:

- C0750480 Count (2)
- C0205173 Duplicate (1)
- ...

Abstract:

- C0750480 Count (2)
- C0205173 Duplicate (3)
- C0007634 Cells (4)
- C1517001 Expected (1)
- C1521828 Rate (3)
- ...

Algorithm



Dataset

- National Library of Medicine's Word Sense Disambiguation (NLM-WSD) Dataset
 - 50 words from the 1998 MEDLINE abstracts
 - 100 instances for each of the 50 words
 - Each instance has been tagged by MetaMap
 - The target word was manually assigned a UMLS concept or None
 - Average number of concepts per ambiguous word is 2.26 (not including None)

Data subsets

- Liu subset
 - Liu, Teller and Friedman 2004
 - 22 out of the 50 words in NLM-WSD
- Leroy subset
 - Leroy and Rindflesch 2005
 - 15 out of the 50 words in NLM-WSD
- Joshi subset
 - Joshi, Pedersen and Maclin 2005
 - 28 out of the 50 words in NLM-WSD
 - (union of Leroy and Liu subsets)

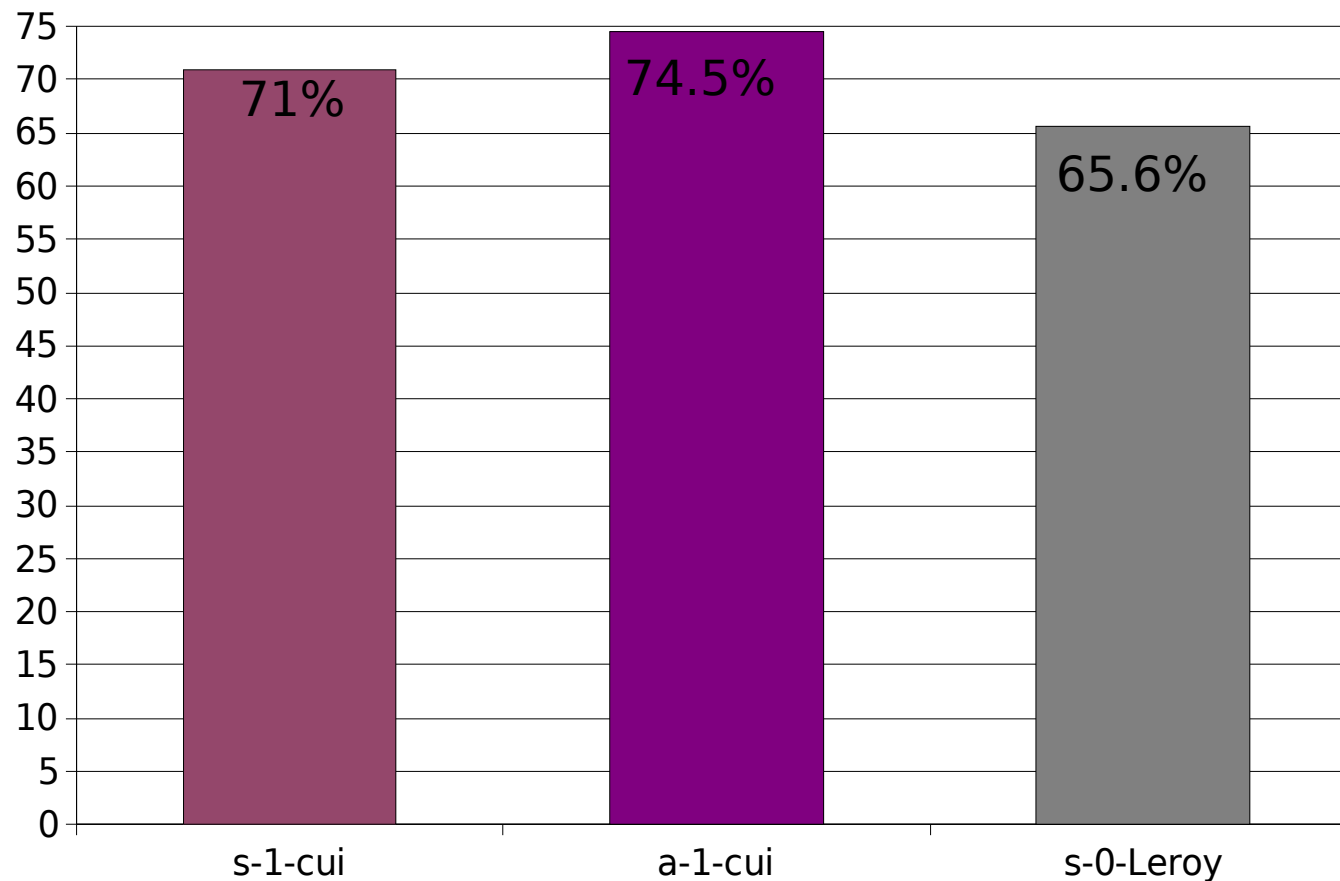
Results

Results for Question 1

Would CUIs be an improvement over semantic types?

Comparative results with Leroy and Rindflesch 2005

Accuracy using Leroy subset



Significance of Differences

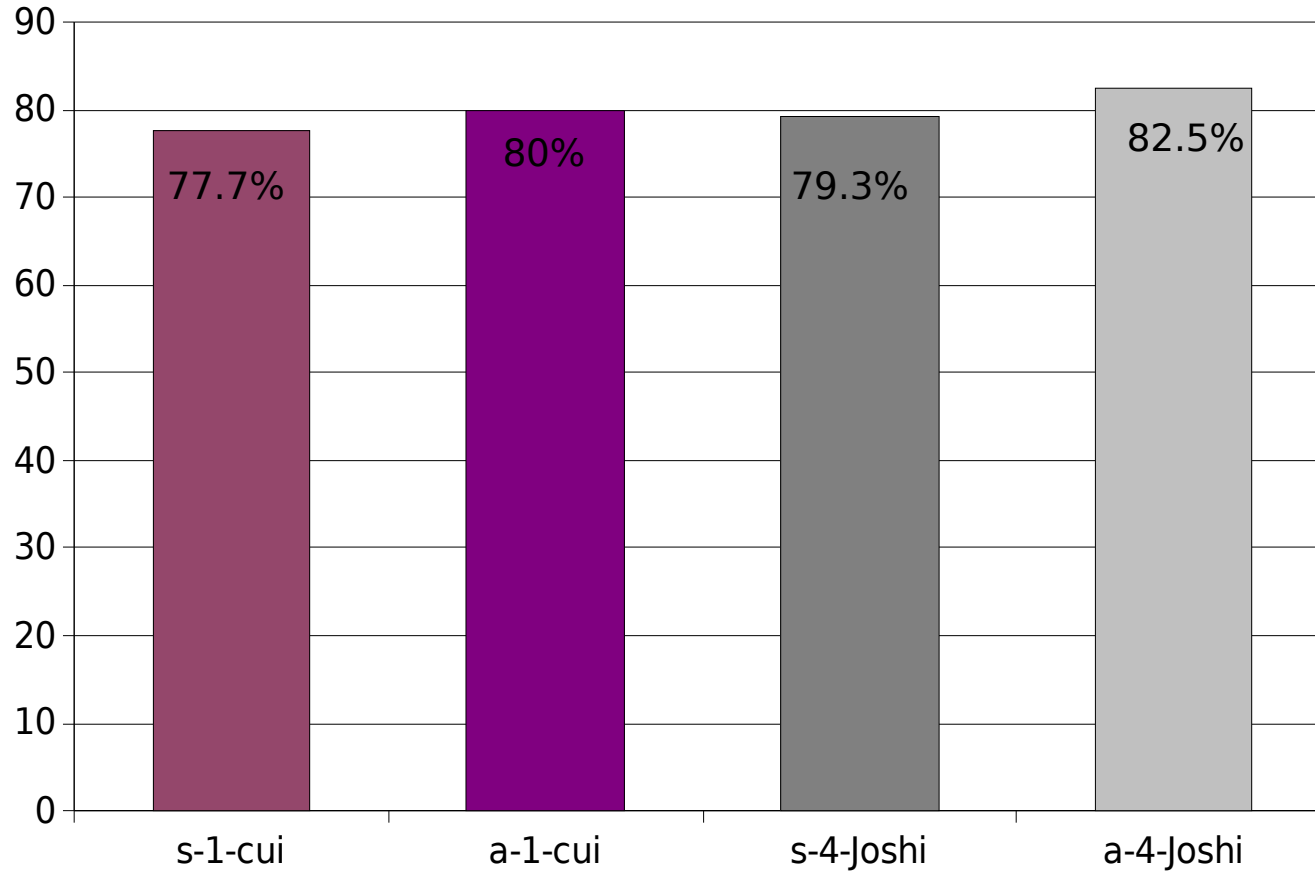
- Pairwise t-test
 - s-1-cui (71%) and s-0-Leroy (65.6%)
 - $p \leq 0.001$
 - a-1-cui (74.5%) and s-0-Leroy (65.6%)
 - $p \leq .00005$

Results for Question 2

Would the biomedical specific feature CUIs be an improvement over the more general feature unigrams?

Comparative results with Joshi, Pedersen and Maclin 2005

Accuracy using Joshi subset



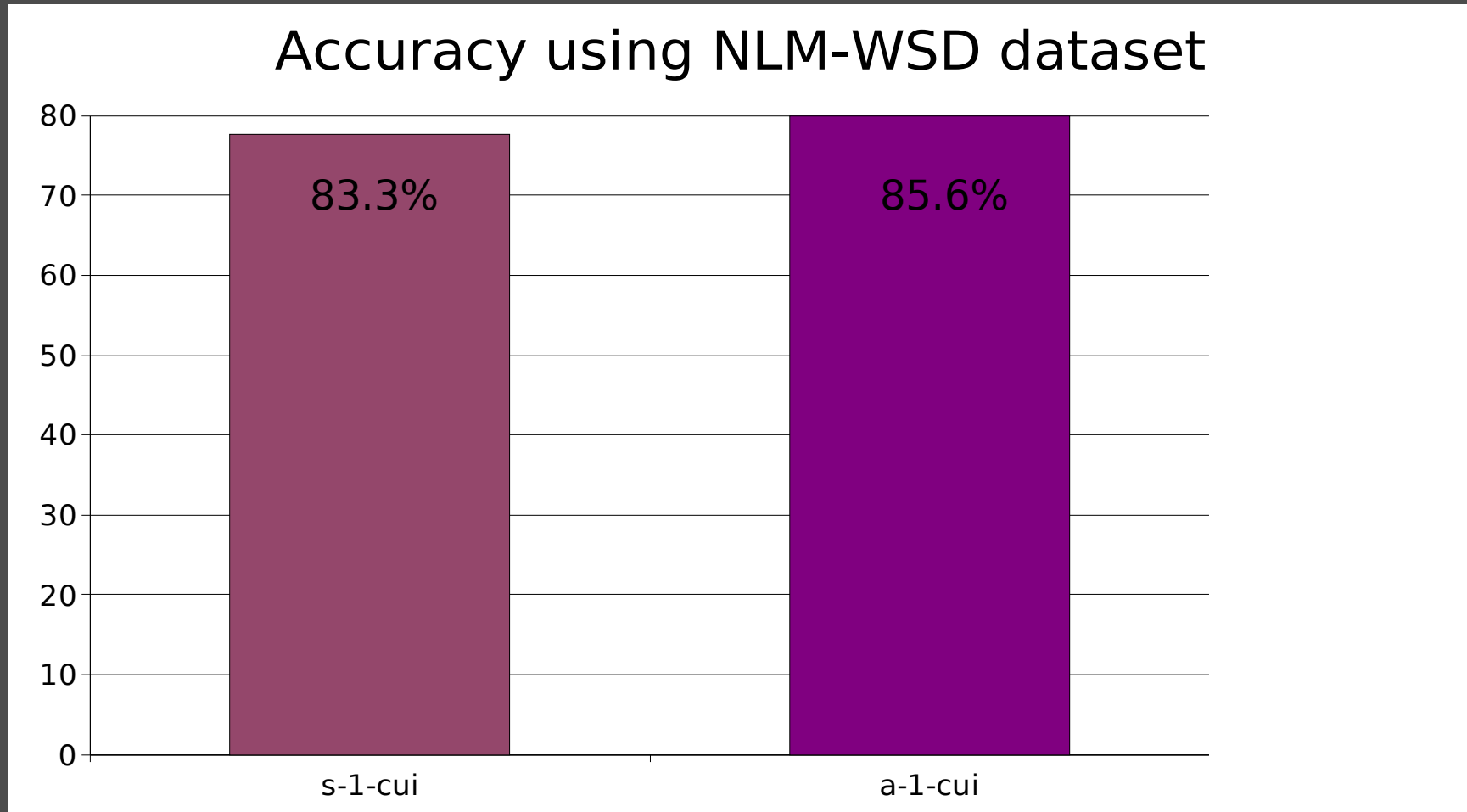
Significance of Results

- Pairwise t-test
 - s-1-cui (77.7%) and s-4-Joshi (79.3%)
 - $p < 0.135$
 - a-1-cui (80.0%) and a-4-Joshi (82.5%)
 - $p < 0.003$

Results for Question 3

Would increasing the size of the context window in which surrounding CUIs are found improve the results, as seen by Joshi, Pedersen and Maclin using unigrams?

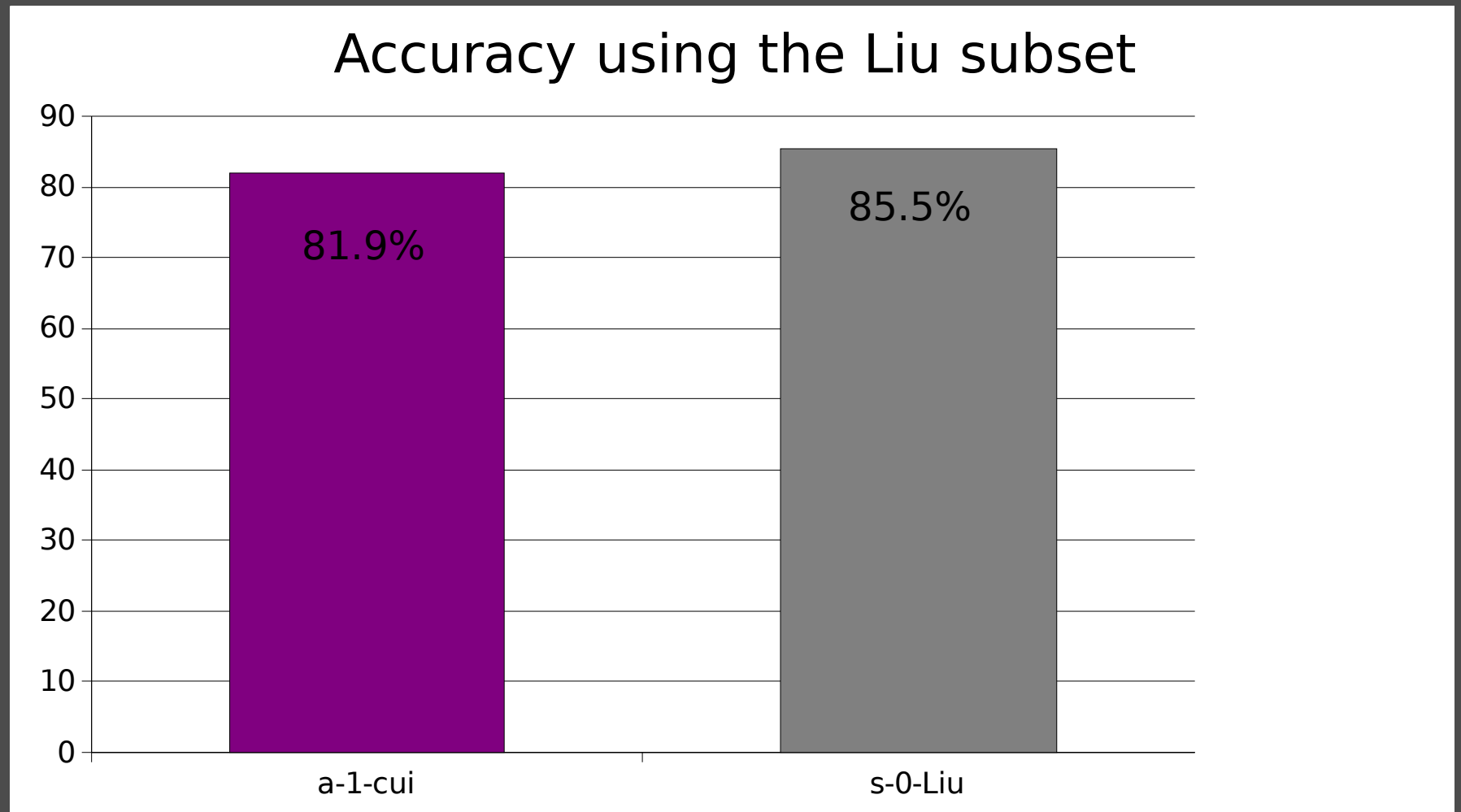
Comparative results between size of context window



Significance of Results

- Pairwise t-test
 - s-1-cui (83.3%) and a-1-cui (85.6%)
 - $p < 0.0006$

Comparative results with Liu, Teller and Friedman 2004



Significance of Results

- Pairwise t-test
 - a-1-cui (81.9%) and s-1-Liu (85.5%)
 - $p < 0.001$

Conclusions

- CUIs result in more accurate disambiguation than semantic types and are comparable to unigrams
- Incorporating more surrounding context improves the results
- MetaMap generates useful information that can be used as features for supervised disambiguation

Future Work

- Combination approach
- Exploring additional UMLS features
- Unsupervised approach using information from the UMLS

Software and Data

- CuiTools version 0.05
 - <http://cuitools.sourceforge.net>
- NLM-WSD Dataset
 - <http://wsd.nlm.nih.gov>
- Pairwise t-test
 - <http://www.quantitativeskills.com/sisa/statistics/>