

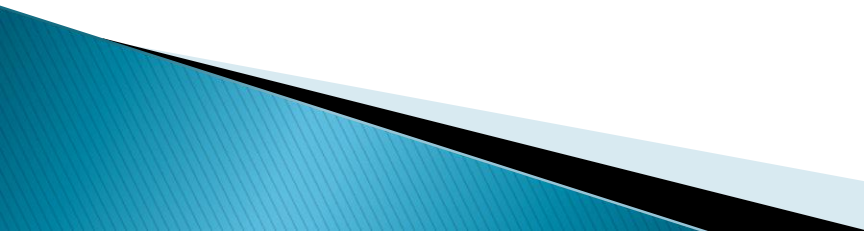
How can a word be disambiguated in a set of documents:
using **recursive Lesk** to select relevant records

Diego Chavarro*, Yuxian Liu

*SPRU, University of Sussex, Brighton, UK



Background

- ▶ Having the right dataset for analysis is crucial in research.
 - ▶ Usually, methodologies start with the analysis phase and do not emphasize sufficiently the importance of data cleaning.
 - ▶ Mistakes in datasets can appear for a variety of reasons (codification, missing values, file corruption, algorithms, etc).
 - ▶ Here we address mistakes produced by the multiple meanings of search terms.
- 

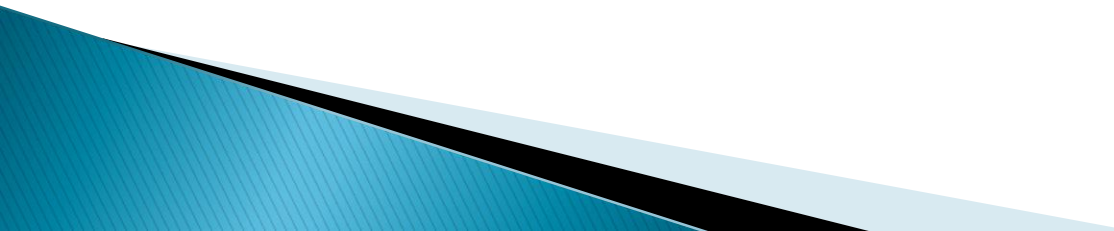
The problem

- ▶ We intent to search and collect bibliographic records for Human Epidermal Growth Factor 2, a biomarker of great importance for cancer detection. Its most used acronym is HER 2, but there are others: HER2/NEU, C-erbB-2, etc.

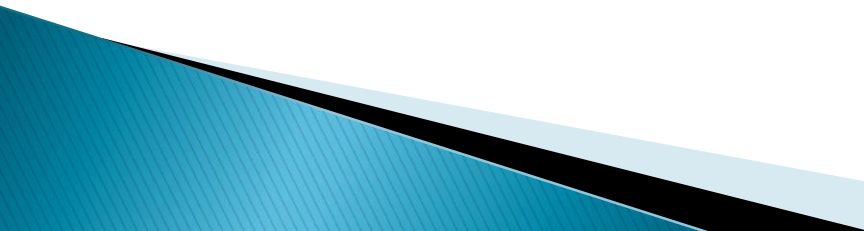
BUT

- ▶ “Her 2” can be used as in “her 2 children” which has nothing to do with Human Epidermal Growth Factor 2. In addition, “children” can be replaced by any noun. Between “her 2” and the noun any adjective can be added.
- ▶ Since the Web of Science (WoS) ignores all punctuations, any punctuation can be added in between.
- ▶ However, the fact that one item has “her 2 children” does not necessarily mean it is not what we need. Some of the articles dealing with Human Epidermal Growth Factor 2 could include the expression “her 2 children”. These configurations make it very difficult for us to formulate an effective search string.

Research Question

- ▶ Is there a way to automatically disambiguate a word to select only the records related to Human Epidermal Growth Factor 2?
 - ▶ How effective it is as compared to manual inspection?
- 

Approach

- ▶ Since the problem of disambiguation can be seen as a decision between several senses of the same string, we use a dictionary to make this decision.
 - ▶ We compare the dictionary's definition of a word with a corpus of research papers.
 - ▶ We then select the papers that are very similar to the definition given by the dictionary.
 - ▶ This is called the Lesk algorithm.
- 

Test dataset

	Search string	Number of items
#1	TS="her 2"	8,542
#2	string 1	26,972
#3	#2 AND #1	6396
#4	#1 NOT #3	2,146

- ▶ Firstly we use TS="her 2" to retrieve the data from the web of science, recalling 8,542 items. Among these items we exclude those definitely related to the bio-marker Epidermal growth factor receptor 2. We then have 2,146 records that we cannot judge if they are related to the bio-marker her 2 or not.

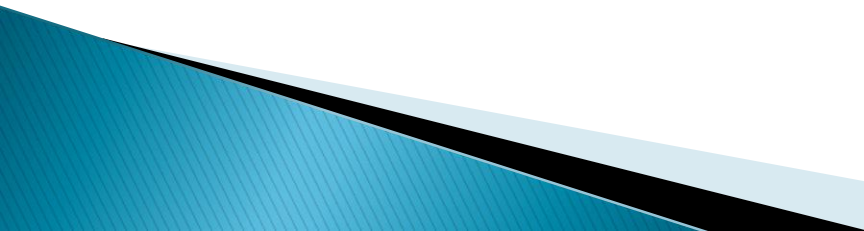
Standard set

- ▶ We tested 2,146 records that contained both related and unrelated records.
- ▶ We proceeded to manually inspect the records and found out

total	Records not related	Records related
2146	98	1873

- ▶ we compared this with our algorithm.

Recursive Lesk algorithm (rLesk)

- ▶ The Lesk algorithm is a classical algorithm for word sense disambiguation introduced by Lesk (1986).
 - ▶ The Lesk algorithm is based on the assumption that words in a given "neighbourhood" tend to share a common topic.
 - ▶ Our idea of a recursive Lesk is to add terms to the dictionary after each string similarity calculation.
 - ▶ In the next iteration we calculate the similarity between the unclassified articles and the new augmented dictionary.
 - ▶ After all calculations, we get two sets: the items that have a high similarity in their topics and the other set that are distant.
- 

The general procedure:

- ▶ 1. load the dictionary with a starting definition of the term (A).
- ▶ 2. load the corpus with the records to classify (B).
- ▶ 3. for each record in B
 - ▶ 3.1. remove stopwords from record B(i).
 - ▶ 3.2. clean string B(i): colon, non printable characters, other non-significant characters.
 - ▶ 3.3. split string B(i) into an array of words
 - ▶ 3.4. For each word in array of words
 - ▶ 3.4.1. check if word is in dictionary
 - ▶ 3.4.2. count number of words that matched dictionary
 - ▶ 3.4.3. calculate percentage of words that matched the dictionary:
number of matched words / number of words in string B(i)
 - ▶ 3.4.4. If percentage \geq threshold, classify record as relevant; Otherwise continue.
- ▶ 4. Add identified records to the dictionary
- ▶ 5. start again, until reaching max. number of iterations.

Definition of the dictionary

- ▶ The lesk algorithm relies on a dictionary composed of a set of words that define a term.
- ▶ The algorithm classifies a string according to its similarity with the dictionary. The algorithm iterates over a corpus and grows the dictionary in each iteration by adding the records that meet a certain percentage threshold of string similarity.
- ▶ To identify the records related to Her 2 as a bio-marker, a paper that has been cited more than 6,000 times was used as a seed.
- ▶ Slamon, Clark, Wong, Levin, Ullrich, and McGuire (1987). Human Breast Cancer: Correlation of Relapse and Survival with Amplification of the HER-2/neu Oncogene. *Science*. 235. 177-181, 235.

results

	algorithm yes	algorithm no
Manually- yes	0.56	0.44
Manually no	0.06	0.94

- ▶ Not very accurate finding the records that are related to the bio-marker her 2.
- ▶ But it is very accurate on the records that are not related to the bio-marker her 2.
- ▶ It is not unexpected, since our algorithm is based on the similarity of topics and there are different aspects that deal with the bio-marker her 2. In addition, we started only with one article as the seed to define the dictionary. So, the records that are picked up by our algorithm form just one aspect of her 2 that is related to the correlation of relapse and survival with amplification of her-2 oncogene. The other records concerning the medicine cannot be picked up by our algorithm, but could be increased if we add more articles to the seed at the first stage.

Further definitions

we use VoSviewer to see how many aspects the research topic her 2 have concerned

We can see there are four clusters: the statues of her2 and the methods to test the statues;

The theraputic and its side-effecttion;

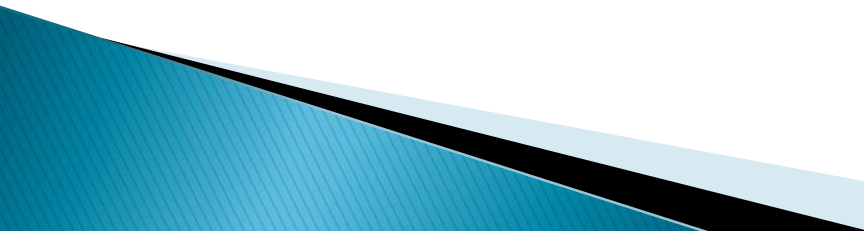
The gene family;

Immunotherapy.


Seeds in the definition of the dictionary on biomarker her 2

- ▶ And then select representative articles from each cluster.
- ▶ **Lapatinib plus Capecitabine for HER2-Positive Advanced Breast Cancer**
- ▶ **Nielsen, T. O. Hsu, F.D., Jensen, K. Cheang, M., Karaca, G. et al (2004). Immunohistochemical and Clinical Characterization of the Basal Like Subtype of Invasive Breast Carcinoma. Clinical Cancer Research. 10, 5367-5347.**
- ▶ **Mellinghoff, I. K. et al (2005). Molecular Determinants of the Response of Glioblastomas to EGFR Kinase Inhibitors. New England Journal of medicine. 353; 19.2012-2014.**
- ▶ **Romond, E.H et al (2005). Trastuzumab plus Adjuvant Chemotherapy for Operable HER2-Positive Breast Cancer. New England Journal of medicine. 19.3159-3167.**
- ▶ **Sørli T. et al (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS. 98(19).10869-10847**

Hypotheses

- ▶ Including different articles in the seed we should pick up all articles in different aspects of a research topic.
 - ▶ We also can scrutinize how these different aspects are related with each other so that we can understand how a research topic is recognized.
 - ▶ After several loops, we have got some words that are related to the research topic. We can use these words as descriptors to specify these words that have general meaning, which will help to achieve high accuracy and completeness.
- 

Future work

- ▶ This is only the first stage in a more comprehensive test of algorithms to disambiguate words in data gathering for research.
 - ▶ Other techniques such as naïve Bayes, neural networks, and support vector machines will be used in later stages to compare their advantages and disadvantages.
 - ▶ We believe that the automation of this process can help to ensure the accuracy of research datasets in situations in which big datasets are used.
 - ▶ Producing a reliable algorithm to disambiguate words will ensure the validity of the data. This is important because the results used as evidence in scientific papers and policy documents are key to make decisions that have an impact on society.
- 

Thanks you for your attention

