Text Mining

- Part of Speech Tagging (POS tagging)
- Named Entity Recognition (NER)
- Relation Extraction

Zhaopeng Xing
Doctoral Student at Carolina Health Informatics Program in Biomedical Health Informatics

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

The Big Picture – Information extraction

Noun,

Nationa

From Wikipedia, t

"NIH" redirec

The **National In**
for biomedical ar
Department of H
conducts its own
research funding

As of 2013, the I
fellows in basic,
as of 2003, the e
billion.[6]

The NIH compris
scientific accomp
disorder, and the

In 2019, the NIH
contributors to pa

Verb, A inc

Agency overview	
Formed	1887
Preceding agency	Hygienic Laboratory
Headquarters	Bethesda, Maryland, U.S.
Employees	20,262 (2012),[1] including 6,000 research scientists (2019).[2]
Annual budget	▲ US$39 billion (2019)[2] ▲ US$37 billion[3] (2018)[4]
Agency executives	Francis Collins, Director Lawrence Tabak, Principal Deputy Director
Parent agency	Department of Health & Human Services
Child agencies	National Cancer Institute National Institute of Allergy and Infectious Diseases National Institute of Diabetes and Digestive and Kidney Diseases National Heart, Lung, and Blood Institute National Library of Medicine
Website	NIH.gov

ıber

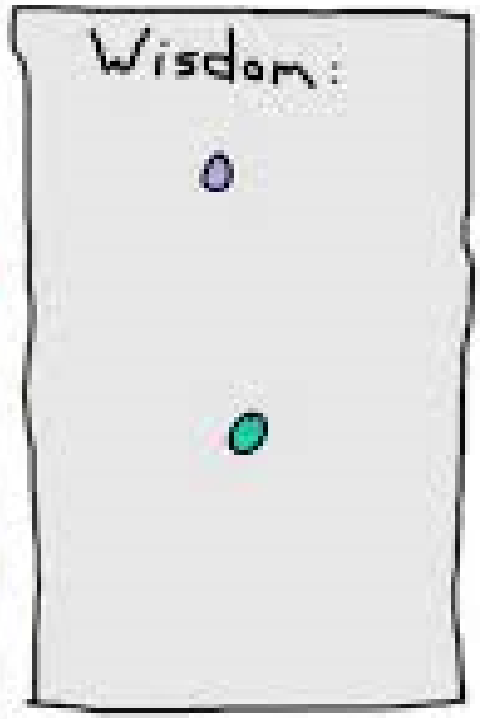
Noun, Organization.
A part of

nited States government responsible
ow part of the United States
d in Bethesda, Maryland. The NIH
and provides major biomedical
gram.

s and more than 4,000 postdoctoral
earch institution in the world,[5] while,
annually in the U.S., or about US$26.4

ines and is responsible for many
y, the use of lithium to manage bipolar
B), and human papillomavirus (HPV).[7]

ndex, which measured the largest
]



Outline

Goal

- What is Part of Speech Tagging (POS tagging) and how it is used in text mining?
- What is Named Entity Recognition (NER) and how it is used in text mining?
- What is the dependency structure and how to extract predicate-argument relation?

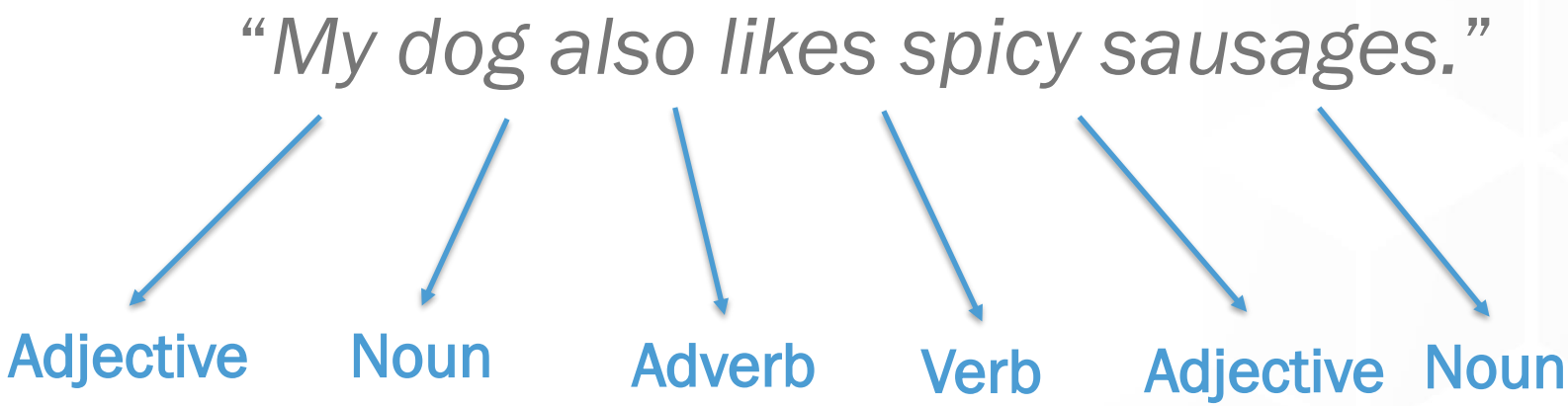
Part of Speech Tagging (POS tagging)

Named Entity Recognition (NER)

Relation Extraction

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

POS tagging


"My dog also likes spicy sausages."
Adjective
Noun
Adverb
Verb
Adjective
Noun

Non-trivial function: Disambiguation

Like

*Please let me know how you would **like** to proceed .* Verb

***Like** most people, I'd prefer to have enough money not to work.* Preposition

Having the same characteristics as...

*She looks **like** she is about to cry .* Conjunction

...as if...

Penn treebank and Universal tag set

Penn treebank

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

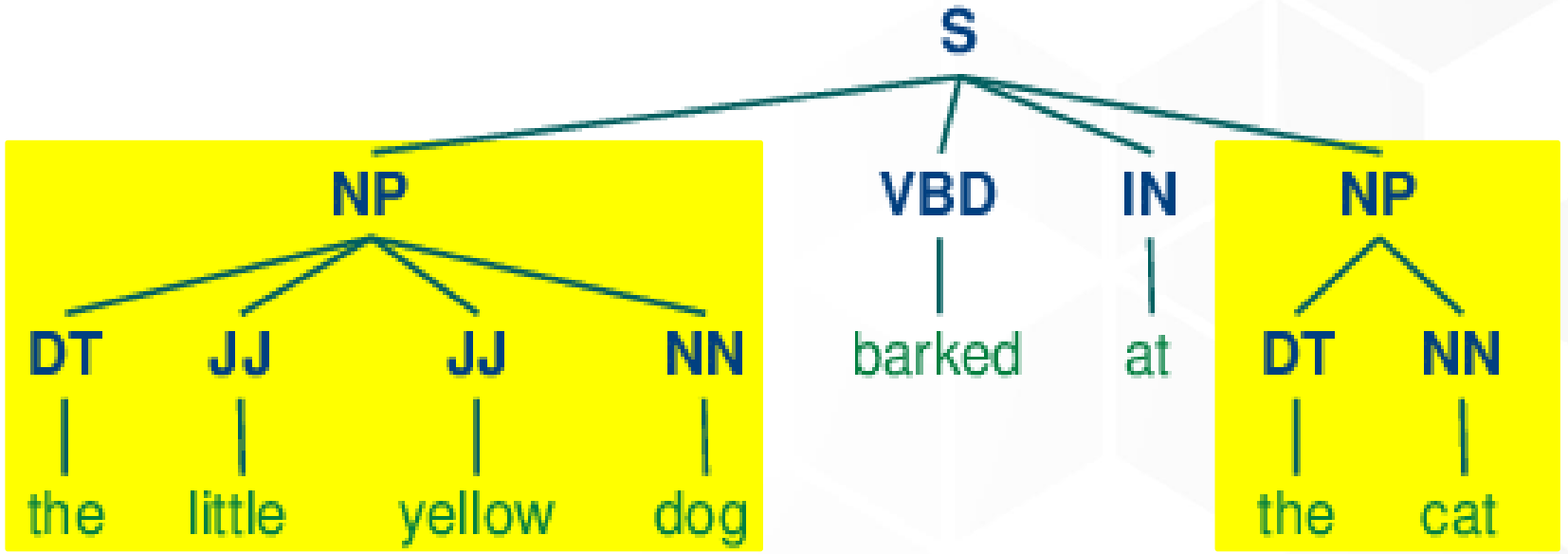
Universal tags

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

POS tags – Where we can use it?

Syntactic chunking

Noun phrase
Verb phrase
Preposition phrase
Adjective phrase
Adverb phrase



Source: https://www.nltk.org/book/ch07.html

POS tags – Where we can use it?

Sentiment analysis

- Adjective and adverb as sentiment indicators

Positive: Good selection of food , fairly reasonable prices. Typical American Chinese food. Waitress was attentive.

Negative: Clearly one of the worst restaurants in the country, not because of their food or service but rather their misleading/fraudulent business practices

Source: https://www.yelp.com/

More to think about

What are some other use cases of POS tagging in information extraction? (Maybe google “POS tagging”

Task 1: POS tags and noun phrase extraction

Notice: Your work in Jupyter notebook will not be saved. Please download it.

Part of Speech Tagging (POS tagging)

Named Entity Recognition (NER)

Relation Extraction

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

NER

Named Entity Recognition

Find and classify information units from text

[Who] did **[What]** at **[When]** in **[Where]**

NER – Example

"The NIH was founded in 1887 and is now part of the United States Department of Health and Human Services. The NIH is located in Maryland, U.S. and has nearly 1,000 scientists and support staff. The NIH obtained US$39 billion from Congress in 2019"

ORG	NIH, the United States Department of Health and Human Services, Congress
DATE	1887, 2019
MONEY	US$39 billion
CARDINAL	Nearly 1,000
GPE	Maryland, U.S.

Biomedical NER

- Patient history, e.g., [Problem], [Test], [Treatment] in biomedical NER

She has had <treatment> physical therapy </treatment> and recovered completely from that .
<test> Initial examination </test> showed <problem> bruising </problem> around the left eye , normal lung examination , normal heart examination , normal neurologic function with a baseline decreased mobility of <problem> her left arm </problem> .

Problem	Bruising, her left arm
Test	Initial examination
Treatment	Physical therapy

More to think about

In the domain you are interested in, what types of entities are important to extract?

Task 2: Extract named entities

Notice: Your work in Jupyter notebook will not be saved. Please download it.

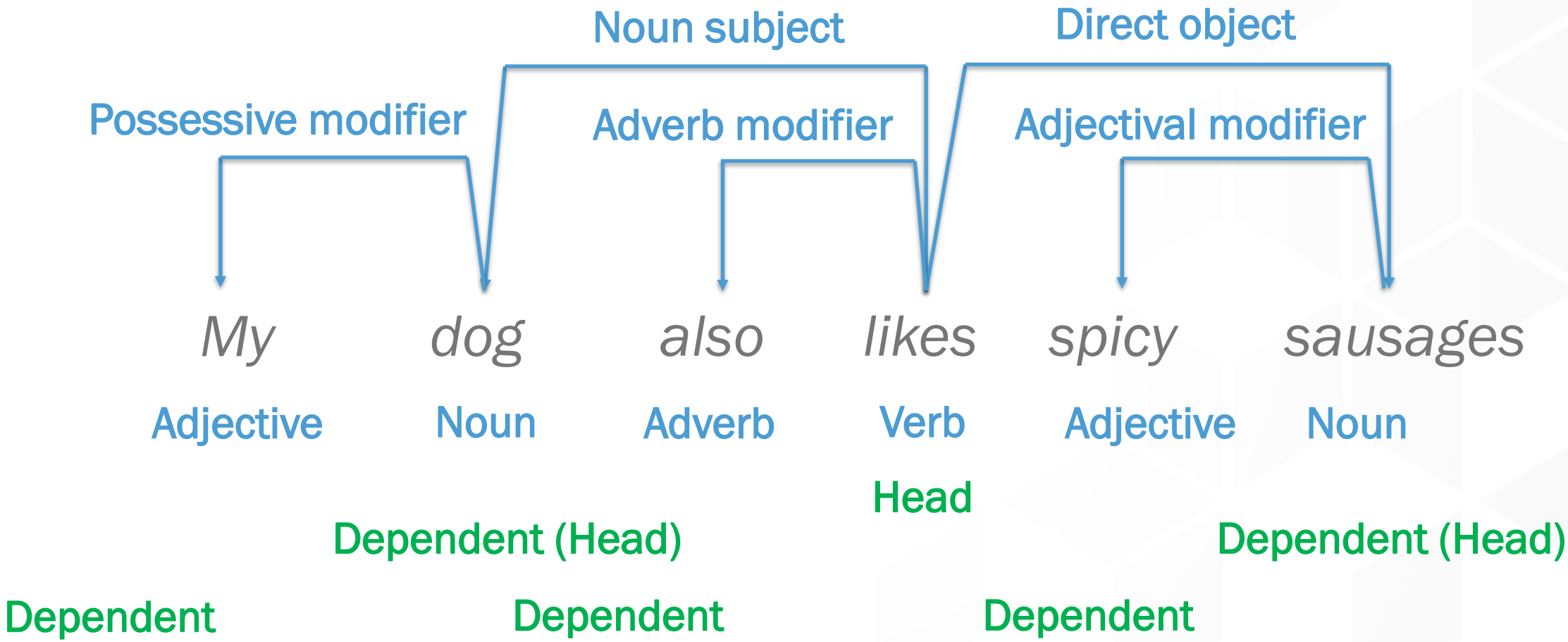
Part of Speech Tagging (POS tagging)

Named Entity Recognition (NER)

Relation Extraction

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Syntactic Dependency Structure



More to think about

Task 3: How about the syntactic dependency structure of the following complex sentence?

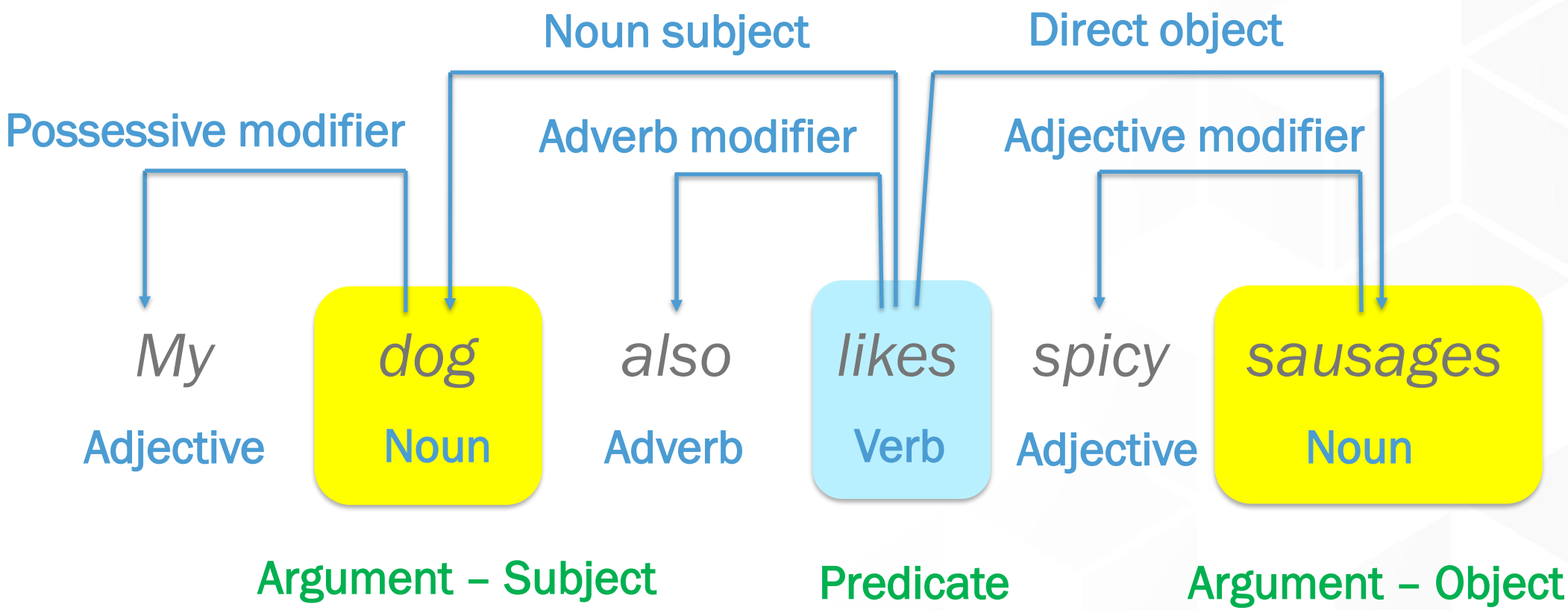
I remember that you have given Tom a gift

Bell makes and distributes computer products.

The NIH is located in Maryland, U.S. and it contains nearly 1,000 scientists and support staff.

Notice: Your work in Jupyter notebook will not be saved. Please download it.

Predicate–argument structure



Biomedical relation extraction

- Evidence-based practice, e.g., Drug efficacy, interaction, side effects

*"We used **hemofiltration** to treat a **patient with digoxin overdose** that was complicated by **refractory hyperkalemia**"*

Subjects	Predicate	Objects
Hemofiltration	TREATS	Patient, digoxin overdose
Hyperkalemia	COMPLICATES	Digoxin overdose
Digoxin overdose	PROCESS OF	Patient

(Lai, Swaminathan, Pun, & Vallance-Owen, 1986)

More to think about

Task 4 (Optional): extract subject-predicate-object relation by analyzing the syntactic dependency

Notice: Your work in Jupyter notebook will not be saved. Please download it.

- You can play with the biomedical relation extraction tool with the text you are interested in or the sample text below.

This study demonstrates that netilmicin is a safe and effective antibiotic that can be used as a first choice treatment of acute bacterial conjunctivitis.

Wrap-up and take-aways

- Part of Speech Tagging
- Named Entity Recognition
- Relation Extraction

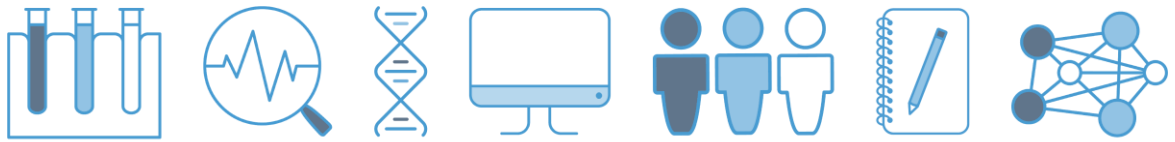
- Unstructured data → structured data
- Information objects and relation

CAROLINA HEALTH INFORMATICS PROGRAM

We are grateful to the following organizations for their support:

UNITED HEALTH FOUNDATION®

NIH National Institutes of Health



Program for Precision Medicine in Health Care

Interested in an MPS or a Ph.D. in digital health?

Drop us a line: chip@unc.edu.

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Reference

Jurafsky, D., & Martin, J. H. (2016). Speech and Language Processing.(3rd ed.)

- Overview:

Task of information extraction, https://people.cs.umass.edu/~mccallum/courses/inlp2007/lect20-ie.ppt.pdf

- POS part:

Fan, J., Prasad, R., Yabut, R. M., Loomis, R. M., Zisook, D. S., Mattison, J. E., & Huang, Y. (2011). Part-of-speech tagging for clinical text: wall or bridge between institutions? *AMIA Annual Symposium Proceedings*, *2011*, 382–391.

- NER:

Ramshaw, L. A., & Marcus, M. P. (1999). Text Chunking Using Transformation-Based Learning. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, & D. Yarowsky (Eds.), *Natural language processing using very large corpora* (Vol. 11, pp. 157–176). Springer Netherlands. https://doi.org/10.1007/978-94-017-2390-9_10

Reference (cont.)

Rindflesch, T.C. and Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics, 36(6):462-477.

Kundeti, S. R., Vijayananda, J., Mujjiga, S., & Kalyan, M. (2016). Clinical named entity recognition: Challenges and opportunities. *2016 IEEE International Conference on Big Data (Big Data)*, 1937–1945. https://doi.org/10.1109/BigData.2016.7840814

Wu, Y., Jiang, M., Xu, J., Zhi, D., & Xu, H. (2017). Clinical named entity recognition using deep learning models. *AMIA Annual Symposium Proceedings*, *2017*, 1812–1819.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., & Liu, H. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, *77*, 34–49. https://doi.org/10.1016/j.jbi.2017.11.011

- Relation extraction:

Nivre, J. (2005). Dependency grammar and dependency parsing. *MSI Report, 5133*, 1–32.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., … Liu, H. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics, 77*, 34–49.

Figure Citation

Text information extraction
https://www.gapingvoid.com/

POS Tagging
http://acl.ldc.upenn.edu/J/J93/J93-
https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.
https://universaldependencies.org/u/pos/index.html

NER
http://text-machine.cs.uml.edu/cliner/samples/cliner-sample-output.pdf

NLP applications in clinical information extraction

Name	Description	Website
cTAKES	Open-source NLP system based on UIMA framework for extraction of information from electronic health records unstructured clinical text	http://ctakes.apache.org/
MetaMap	National Institutes of Health (NIH)-developed NLP tool that maps biomedical text to UMLS concepts	https://metamap.nlm.nih.gov/
MedLEE	NLP system that extracts, structures, and encodes clinical information from narrative clinical notes	http://zellig.cpmc.columbia.edu/medlee/
KnowledgeMap Concept Indexer (KMCI)	NLP system that identifies biomedical concepts and maps them to UMLS concepts	https://medschool.vanderbilt.edu/cpm/center-precision-medicine-blog/kmci-knowledgemap-concept-indexer
HITEx	Open-source NLP tool built on top of the GATE framework for various tasks such as principal diagnoses extraction and smoking status extraction	https://www.i2b2.org/software/projects/hitex/hitex_manual.html
MedEx	NLP tool used to recognize drug names, dose, route, and frequency from free-text clinical records	https://medschool.vanderbilt.edu/cpm/center-precision-medicine-blog/medex-tool-finding-medication-information
MedTagger	Open-source NLP pipeline based on UIMA framework for indexing based on dictionaries, information extraction, and machine learning–based named entity recognition from clinical text	http://ohnlp.org/index.php/MedTagger
ARC	Automated retrieval console (ARC) is an open-source NLP pipeline that converts unstructured text to structured data such as Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) or UMLS codes	http://blulab.chpc.utah.edu/content/arc-automated-retrieval-console
Medtex	Clinical NLP software that extracts meaningful information from narrative text to facilitate clinical staff in decision-making process	https://aehrc.com/research/projects/medical-free-text-retrieval-and-analytics/#medtex
CLAMP	NLP software system based on UIMA framework for clinical language annotation, modeling, processing and machine learning	https://sbmi.uth.edu/ccb/resources/clamp.htm
MedXN	A tool to extract comprehensive medication information from clinical narratives and normalize it to RxNorm	http://ohnlp.org/index.php/MedXN
MedTime	A tool to extract temporal information from clinical narratives and normalize it to the TIMEX3 standard	http://ohnlp.org/index.php/MedTime
PredMED	NLP application developed by IBM to extract full prescriptions from narrative clinical notes	

(Wang et al., 2018)

Text Mining

- Part of Speech Tagging (POS tagging)
- Named Entity Recognition (NER)
- Relation Extraction

Zhaopeng Xing
Doctoral Student at Carolina
Health Informatics Program in
Biomedical Health Informatics

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL