

ΓΕ77
COMPUTATIONAL LINGUISTICS

Athanasios N. Karasimos

akarasimos@gmail.com

BA in Linguistics | National and Kapodistrian University of Athens

Lecture 12 | Wed 6 Jun 2018

ADVANCED TOPICS OF COMPUTATIONAL LINGUISTICS

PART-OF-SPEECH TAGGING

PART-OF-SPEECH TAGGING

- **Grammatical categories: parts-of-speech**
 - Nouns: people, animals, concepts, things
 - Verbs: expresses action in the sentence
 - Adjectives: describe properties of nouns

>> The Substitution test

Mary _____ the chicken.

THE PART-OF-SPEECH TAGGING TASK

Input: **the lead paint is unsafe**

Output: **the/Det lead/N paint/N is/V unsafe/Adj**

- Uses:
 - text-to-speech (how do we pronounce “process”?)
 - can differentiate word senses that involve part of speech differences (what is the meaning of “deal”)
 - can write regexps like `Det Adj* N*` over the output (for filtering collocations)
 - can be used as simpler “backoff” context in various Markov models when too little is known about a particular history based on words instead.
 - preprocessing to speed up parser (but a little dangerous)
 - tagged text helps linguists find interesting syntactic constructions in texts (“ssh” used as a verb)

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP\$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP\$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Brown/Penn Treebank tags

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	<i>and, but, or</i>	SYM	Symbol	<i>+, %, &</i>
CD	Cardinal number	<i>one, two, three</i>	TO	"to"	<i>to</i>
DT	Determiner	<i>a, the</i>	UH	Interjection	<i>ah, oops</i>
EX	Existential 'there'	<i>there</i>	VB	Verb, base form	<i>eat</i>
FW	Foreign word	<i>mea culpa</i>	VBD	Verb, past tense	<i>ate</i>
IN	Preposition/sub-conj	<i>of, in, by</i>	VBG	Verb, gerund	<i>eating</i>
JJ	Adjective	<i>yellow</i>	VBN	Verb, past participle	<i>eaten</i>
JJR	Adj., comparative	<i>bigger</i>	VBP	Verb, non-3sg pres	<i>eat</i>
JJS	Adj., superlative	<i>wildest</i>	VBZ	Verb, 3sg pres	<i>eats</i>
LS	List item marker	<i>1, 2, One</i>	WDT	Wh-determiner	<i>which, that</i>
MD	Modal	<i>can, should</i>	WP	Wh-pronoun	<i>what, who</i>
NN	Noun, sing. or mass	<i>llama</i>	WP\$	Possessive wh-	<i>whose</i>
NNS	Noun, plural	<i>llamas</i>	WRB	Wh-adverb	<i>how, where</i>
NNP	Proper noun, singular	<i>IBM</i>	\$	Dollar sign	<i>\$</i>
NNPS	Proper noun, plural	<i>Carolinas</i>	#	Pound sign	<i>#</i>
PDT	Predeterminer	<i>all, both</i>	"	Left quote	<i>(' or ")</i>
POS	Possessive ending	<i>'s</i>	"	Right quote	<i>(' or ")</i>
PP	Personal pronoun	<i>I, you, he</i>	(Left parenthesis	<i>([({ <</i>
PP\$	Possessive pronoun	<i>your, one's</i>)	Right parenthesis	<i>([. } ></i>
RB	Adverb	<i>quickly, never</i>	,	Comma	<i>,</i>
RBR	Adverb, comparative	<i>faster</i>	.	Sentence-final punc	<i>(. ! ?)</i>
RBS	Adverb, superlative	<i>fastest</i>	:	Mid-sentence punc	<i>(; ... --)</i>
RP	Particle	<i>up, off</i>			

POS tag	Tag Name
N	Noun
PREP	Preposition
PUNC	Punctuation
AJ	Adjective
V	Verb
CON	Conjunction
NUM	Number
PRO	Pronoun
DET	Determiner
ADV	Adverb
POSTP	Postposition
RES	Residual
CL	Classifier
INT	Interjection

TAGGED DATA SETS

TAGGED DATA SETSV

- Brown Corpus ([Tag set](#))
 - Designed to be a representative sample from 1961 news, poetry, ...
 - 87 different tags
- Claws5 “C5” vs. Claws7 ([tag set](#))
 - 62 different tags
- Penn Treebank ([tag set](#))
 - 45 different tags
 - Most widely used currently

PART-OF-SPEECH EXAMPLES

• Adjective	JJ	happy, bad
• Adjective, comparative	JJR	happier, worse
• Adjective, cardinal number	CD	3, fifteen
• Adverb	RB	often, particularly
• Conjunction, coordination	CC	and, or
• Conjunction, subordinating	IN	although, when
• Determiner	DT	this, each, other, the, a, some
• Determiner, postdeterminer	JJ	many, same
• Noun	NN	aircraft, data
• Noun, plural	NNS	women, books
• Noun, proper, singular	NNP	London, Michael
• Noun, proper, plural	NNPS	Australians, Methodists
• Pronoun, personal	PRP	you, we, she, it
• Pronoun, question	WP	who, whoever
• Verb, base present form	VBP	take, live

TAGGED SETS: OPEN AND CLOSED

- Closed Set tags
 - Determiners
 - Prepositions
 - ...
- Open Set tags
 - Noun
 - Verb

POS TAGGING: TASK

- Late home after a night out, a youngster attempted to climb into his home down the chimney. He did not want to wake other residents in the Judson Center social services agency; also he had broken his curfew and wanted no trouble.
- In best Santa Claus mode he climbed onto the roof and let himself down the chimney; unfortunately he was too large, and he became stuck. The 17 year old began moaning and was heard and rescued. Fire fighters and police officers from the City of Royal Oak, Michigan, USA, had to pull him out. The youth suffered from minor scrapes and bruises.

WHY IS THIS SUCH A BIG PART OF NLP?

Input: **the lead paint is unsafe**

Output: **the/Det lead/N paint/N is/V unsafe/Adj**

- The first statistical NLP task
- Been done to death by different methods
- Easy to evaluate (how many tags are correct?)
- Canonical finite-state task
 - Can be done well with methods that look at local context
 - (Though should “really” do it by parsing!)

PART OF SPEECH AMBIGUITIES

		VB				
	VBZ	VBZ	VBZ			
NNP	NNS	NNS	NNS	CD	NN	
Fed	raises	interest	rates	0.5	%	in effort to control inflation

DEGREE OF SUPERVISION

- **Supervised:** Training corpus is tagged by humans
- **Unsupervised:** Training corpus isn't tagged
- **Partly supervised:** E.g. Training corpus isn't tagged, but you have a dictionary giving possible tags for each word

CURRENT PERFORMANCE OF TAGGERS

Input: **the lead paint is unsafe**

Output: **the/Det lead/N paint/N is/V unsafe/Adj**

Using state-of-the-art automated method,

- how many tags are correct?
 - About 97% currently
 - But baseline is already 90%
- Baseline is performance of simplest possible method:
- Tag every word with its most frequent tag
- Tag unknown words as nouns

RECIPE FOR SOLVING AN NLP TASK

Input: **the lead paint is unsafe**

Observations

Output: **the/Det lead/N paint/N is/V unsafe/Adj**

Tags

- 1) Data: Notation, representation
- 2) Problem: Write down the problem in notation
- 3) Model: Make some assumptions, define a parametric model (often generative model of the data)
- 4) Inference: How to search through possible answers to find the best one
- 5) Learning: How to estimate parameters
- 6) Implementation: Engineering considerations for an efficient implementation

INFORMATION EXTRACTION

INFORMATION EXTRACTION: DEFINITION

The automatic extraction of information

- Input: possibly limited set of documents
- Output: usually a task-defined template to fill in
- Definitions:
 - Typically idiosyncratically defined for task
 - Can include technology (Semantic Role Labelling, etc.) that helps IE
- Comparison with Question Answering
 - QA more opened ended – depends on questions
 - QA: paragraph output vs. IE structured output
 - Similar techniques

SAMPLES IE TASKS

- Extract instances of people starting jobs and ending jobs
 - Identify: person, start or stop time, company
- Extract instances of Entity1 attacking Entity2, where entities include people, organizations, facilities or vehicles
 - Identify: aggressor, victim, weapon, time
- Extract instances of disease outbreak
 - Identify: victims, disease, start time, time span, location
- Extract advertisements for cameras
 - Identify: seller, brand, model, price, date
- Identify family, social and business relations between individuals

NAMED ENTITY IDENTIFICATION

- Tend to be phrases consisting of proper nouns
 - Capitalization, uniquely identify entity in real world, ...
 - ***The Association for Computational Linguistics***
- Internal structure may differ from common NPs
 - ***Athanasios Karasimos***
- Only certain types are marked
 - Task-specific
 - ACE task: GPE, Person, Organization, Location, Facility
 - In some versions: Vehicle and Weapon

INFOEXTRACTOR TASK

- <http://nlp.stanford.edu:8080/ner/>
- <https://dandelion.eu/semantic-text/entity-extraction-demo/>
- Test the same text and compare them.

AUTOMATIC CONTENT EXTRACTION

- An Entity = a list of coreferential NPs
 - Each of these NPs is a “mention” of the entity
 - Finding coreference will be part of a different lecture
- Types of mentions: names, common nouns, pronouns
- Names: what we have been calling named entities
- Nominal mentions: phrases headed by common nouns
 - same semantic classes: GPE, ORGANIZATION, ...
 - ***that country, the government, the agency, the whimsical***
 - ***pediatrician, the terrorist wearing a hat***
- Pronominal mentions: pronouns
 - Must refer to markable semantic class (e.g., by coreference)
 - ***He, she, it, they, themselves, their, her, everyone, ...***

ACE RELATIONS AND EVENTS

- Predicate + Arguments
- Annotation of Predicate triggers
 - Event mention triggers: words
- Specs discuss choice of nouns/verbs: **launch an attack**
 - Relation mention triggers: grammatical constructions
- ACE specs refer to these constructions as relation classes
- ML must learn which words trigger which relations
- Arguments of Event and Relation Mentions
 - Usually, NPs belonging to ACE Entity classes:
- Named Entities, common noun phrases, pronouns
 - Values – times, extents, crimes, ...
 - http://projects.ldc.upenn.edu/ace/docs/English-Values-Guidelines_v1.2.3.doc
 - Relations always take exactly 2 arguments
 - Event arguments vary in number (and a given argument may be absent)

DETECTING ACE ENTITY MENTIONS

- Detecting ACE common noun mentions:
 - Find common nouns from training corpus
 - Generalize
- Stemming
- WordNet, clustering, or a list of words
 - Identify non-generic cases
 - ***Gardners are lousy plumbers.*** [Generic]
 - ***The gardner was a lousy plumber.*** [Non-Generic]
- Baseline: definite determiners plus past tense → non-generic
- Pronoun Mention – dependent on coreference techniques
- Coreference Component – described in future lecture

ACE RELATIONS

- Relation Entity: set of coreferential relation mentions
 - Same arguments
 - Refer to same predication
- Relation types
 - Physical: Location and Near
 - Part-Whole: Geographical and Subsidiary
 - Per-Social: Business, Family, Lasting-Personal
 - Org-Affiliation: Employee, Owner, Member, ...
 - Agent-Artifact: User-Owner-Inventor-Manufacturer
 - Gen-Affiliation: Citizen-Resident-Religion-Ethnicity, Org-Location-Origin
- Relation Classes: Syntactic environments (sentence internal only)
 - Verbal, Possessive, PreMod, Coordination, Preposition, Formulaic, Participial, Other

ACE RELATION EXAMPLES

- **George Bush traveled to France on Thursday for a summit.**
 - Physical.located(**George Bush, France**)
- **Microsoft's chief scientist**
 - Org-Aff.employment(**Microsoft's chief scientist, Microsoft**)
- **New York police**
 - Part-Whole.Subsidiary(**New York police, New York**)
- **Dick Cheney and a hunting partner**
 - Per-Social.Lasting(**Dick Cheney, a hunting partner**)
- **A linguist from New York**
 - Gen-Aff.CRRE(**A linguist from New York, New York**)

By Adam Mayers

ACE EVENTS

- Event Entity: set of coreferential entity mentions
 - Nonconflicting arguments
- A mention may include a subset of the arguments
 - Refer to same predication (event, state, etc.)
- Event types
 - Life: be-born, marry, divorce, injure, die
 - Movement: transport
 - Transaction: transfer-ownership, transfer-money
 - Business: start-org, end-org, merge-org, declare-bankruptcy
 - Conflict: attack, demonstrate
 - Contact: meet, phone-write
 - Personnel: start-position, end-position, nominate, elect
 - Justice: arrest-jail, release-parole, sue, appeal, pardon, ...

ACE EVENT EXAMPLE

- ***On Thursday, Pippi sailed the ship from Sweden to the South Seas***
 - ANCHOR = sailed
 - ARTIFACT-ARG = Pippi
 - VEHICLE-ARG = the ship
 - ORIGIN-ARG = Sweden
 - DESTINATION-ARG = the South Seas
 - TIME-ARG = Thursday
- Similar to Semantic Role Labeling, but limited to several Frames
 - Like FrameNet
 - fewer frames
 - annotation-based instead of lexicon based
 - Targeted towards specific tasks (unlike PropBank/NomBank)

TIME

- Timex
 - Identifying Absolute Time Expressions
- Regularization
 - Relative Time Expressions
- Regularization
- Relation to document time
- TimeML – temporal relations between 2 args
 - Event and Time [Event ≈ACE Event Mention]
- Event is before/after/at/during/.... Time
 - Event1 and Event2
- Time(Event1) is before/after/at/during/.... Time(Event2)

TIMEX

- Identifies several types of time expressions in text
 - Absolute Time (January 3, 2011 at 5:00 AM)
 - Relative Time (last Thursday)
 - Duration (5 days)
- 2 Types of Markup (XML)
 - Inline:
 - `<TIMEX3 tid="t18" type="DATE" temporalFunction="true" functionInDocument="NONE" value="1990-01-02" anchorTimeID="t17">Jan. 2</TIMEX3>`
 - Offset: `<TIMEX3... start="2015" end="2021"/>`
 - Other than start and end, all the same features

QUESTION- ANSWERING SYSTEM

QUESTION ANSWERING

- Question answering seeks the token or phrase (or passage, document, document set) that is the exact answer to a question
- Questions have many flavors
- Most research is focused on fact questions
- Answers are assumed to be tokens or phrases
- Complete answers are assumed to be found in a single source

TYPES OF QUESTIONS

- Fact :Who killed J.F.F. Kennedy?
- Task : How do I apply for a passport?
- Opinion :What was the best movie this year?
- Definition :Who is Noam Chomsky?
- List :What movies was Bruce Willis in?
- Explanation :What was the cause of the World War II?
- Yes-No : Is it legal to turn right on red in Iowa?

FACT QUESTION EXAMPLES

Q: *When was Mozart born?*

A: 1756

Q: *What is a nanometer?*

A: a billionth of a meter

A: a millionth of a millimeter

Q: *When was The Great Depression?*

A: 1930's

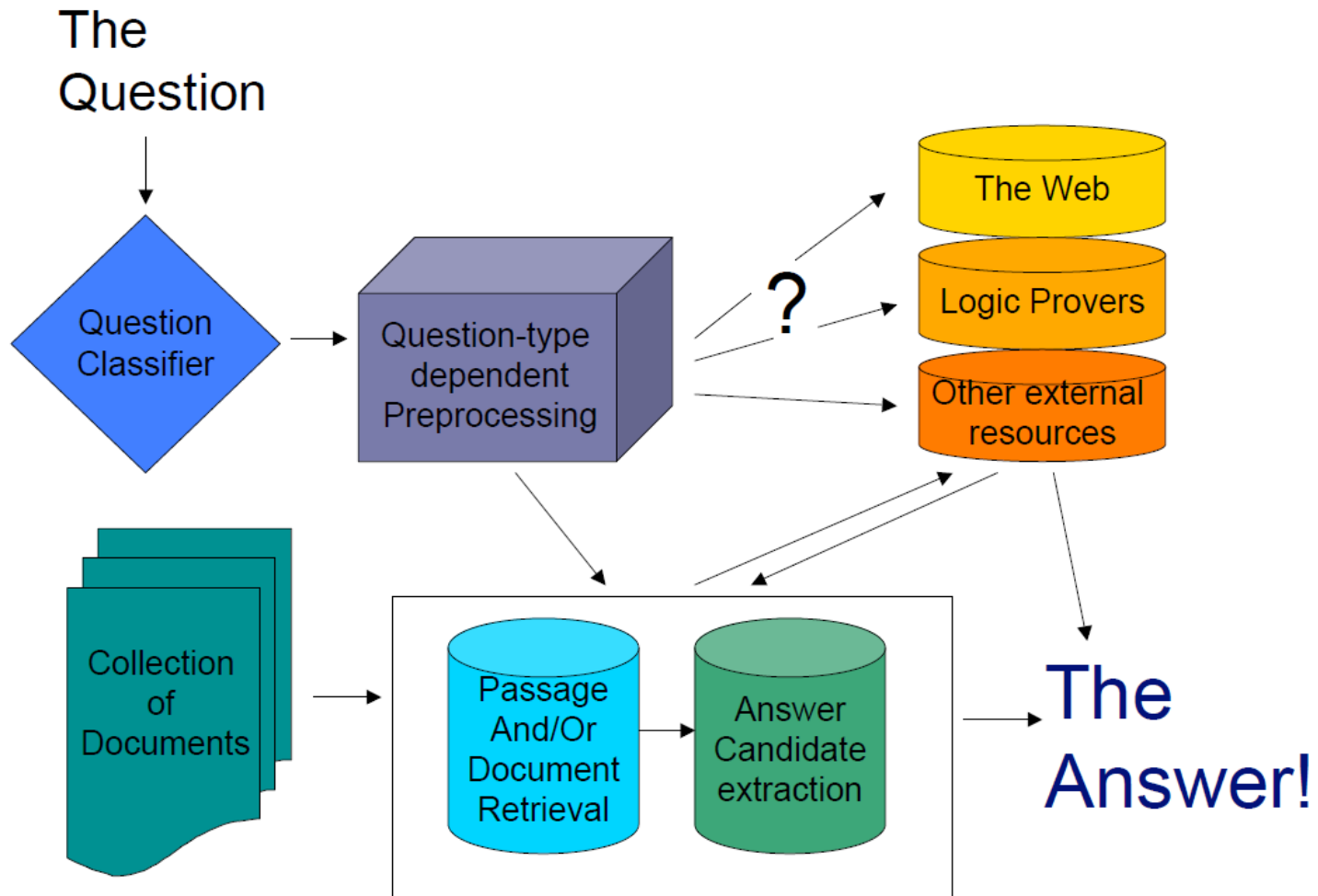
A: 1931

A: 1932

Q: *Who is Absalom?*

A: African-American leader, first black whaling ship captain, desegregated Nantucket's school system.

A: Son of (biblical) David, who betrayed his father



TYPICAL QUESTION ANSWERING SYSTEM

FACT QUESTION CLASSIFICATION

Basically two approaches:

- Classification
 - Advantage: easy to understand/implement, reliable
 - Disadvantage: Doesn't give information other than class
- Regular expressions
 - Advantage: can give information in addition to class
 - Disadvantage: very brittle
 - Classifier Features: POS tags, words, NE tags, WordNet, wh-words, parse trees etc.
- Regular expressions:
 - Simple: wh-words
 - Complex: QA "typology"

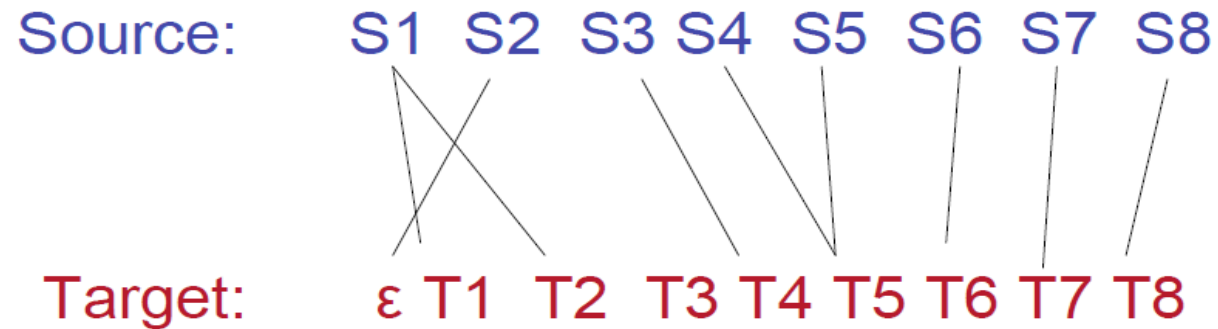
State of the Art: ~90% accuracy

QUESTION CLASSIFICATION: REGEX

- By wh-words, and regular expressions:
 - “Who” => person (Organization? GPE?)
 - “When” => date (Year? Season? Holiday?)
 - “Where” => location (GPE? Organization?)
 - “How”
- “How many” => cardinal number
- “How do” => task question
- “Question typology” extensive regex’s from patterns

IR FOR QA: MODELS

- If we did a good job of IR, QA would be easy
- Passage/sentence retrieval is not just short document retrieval
- Vector Space model
- Query Likelihood
- Translation Models



Translation Models, QA

Source: How high is Mt. Everest?

Target: ε Mt Everest has an elevation of approximately 27,000 ft.

Translation Models train on a parallel corpus, in our case a set of questions and sentences containing answers.

QUESTION TYPE DEPENDENT PROCESSING

- Question rewrites for the web
 - Turn the question into a query, combine multiple evidence
- Logic provers
 - Attempt to reason about the question
- Answer filtering
 - Rule out answers that look right, but can't be
- Question analysis for patterns
 - Patterns in the question suggest patterns in the answer

EXTERNAL RESOURCES

- The web (problematic)
 - Web summaries
 - Answer validation
 - Increasing training data
- POS taggers, NE extractors, noun-phrase chunkers
- Gazetteers, Ontologies, thesauri
 - WordNet, ConceptNet
- Logic Provers
- Previously answered questions

ANSWER EXTRACTION: THE SIMPLE VERSION

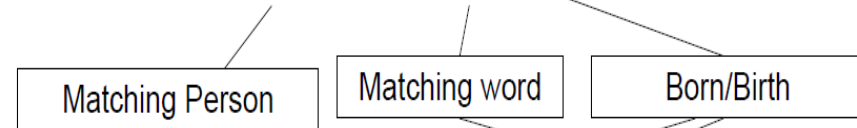
- Extract the answer token that is the correct named entity type from top sentence.
- Extract the answer tokens from top N sentences, and vote.
- Extract answer tokens candidates from top N sentences, validate on the Web

ANSWER TAGGING

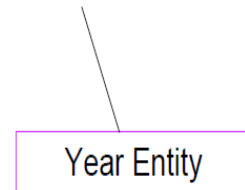
- Treat answer tagging as named-entity tagging
- Answers are frequently not a named entity type (ex. why-famous questions)
- Answer tokens are not predictable and do not always have predictable indicators
- Features of answer tokens are not directly sequential and are often long-range
- Features of one question type may not generalize to other question types

ANSWER TAGGING (EASY)

Q: When were Shakespeare's twins born ?



A: Two years later came the birth of the Shakespeare twins Judith and Hamnet , girl and boy , baptised in 1585 .



- Determine the answer type of the question
- Retrieve a good sentence
- Return the appropriate named entity

ANSWER TAGGING (HARDER)

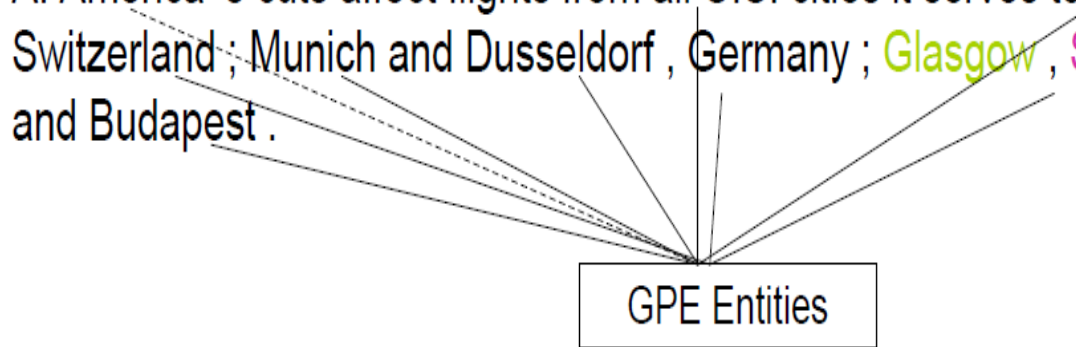
- Determine the answer type of the question
- Retrieve a good sentence
- Return the appropriate named entity

Q: Where is **Glasgow** ?

A: The recession came late to **Glasgow** , as it did to the rest of **Scotland**.

A: America 's cuts affect flights from all U.S. cities it serves to Zurich , Switzerland ; Munich and Dusseldorf , Germany ; **Glasgow** , **Scotland** and Budapest .

GPE Entities



SOME THINGS TO CONSIDER...

- For any given question type, there are potentially hundreds of ways to express the answer.
- Learning patterns depends on multiple unique examples of the same pattern.
- Newswire data has a limited number of examples of any given pattern.
- Newswire data is repetitive: there are many identical examples with different doc ids.

READINGS

