CS-5340/6340 Fall 2017 Project Description

The project for this class will be to design and build an information extraction (IE) system for Latin American terrorism news articles. You will work in a two-person team to build an event extraction system from scratch! Your program should process news stories about terrorism and extract several types of information. A sample news story is shown below:

> DEV-MUC3-0644 (GE)
>
> BOGOTA, 7 NOV 89 (AFP) – [TEXT] SIX PEOPLE WERE KILLED AND FIVE WOUNDED TODAY IN A BOMB ATTACK THAT DESTROYED A PEASANT HOME IN THE TOWN OF QUINCHIA, ABOUT 300 KM WEST OF BOGOTA, IN THE COFFEE-GROWING DEPARTMENT OF RISARALDA, QUINCHIA MAYOR SAUL BOTERO HAS REPORTED.
>
> A SPOKESMAN FOR THE 8TH ARMY BRIGADE SAID THAT THE ATTACK WAS APPARENTLY THE WORK OF GUERRILLAS OF THE MAOIST POPULAR LIBERATION ARMY, WHO SUPPOSEDLY PLANTED THE BOMB IN THE HOME OF ANTONIO ZULUAGA, LOCATED AT VEREDA LA UNION, A FEW KILOMETERS FROM QUINCHIA.
>
> BOTERO SAID THAT THE MOTIVE FOR THE BOMBING IS UNKNOWN.

Each news story describes exactly one terrorist event. Your IE system should produce one *template* per story containing information about the terrorist event. Each template will contain 7 slots:

ID:	News story identifier (e.g., DEV-MUC3-0644)
INCIDENT:	The type of event. There are 5 possible values: {*arson, attack, bombing, kidnapping, robbery*}
WEAPON:	weapons used in the event
PERP INDIV:	individuals who perpetrated the event
PERP ORG:	organizations believed to be responsible for the event
TARGET:	physical targets of the event
VICTIM:	victims of the event

The ID slot should contain the identifier of the news story. The INCIDENT slot should contain a keyword corresponding to one of the 5 event types. All of the other slots are *string slots*, which should contain arbitrary strings extracted exactly from the news story (typos and all! :). Each string slot may contain multiple items, which should appear on separate lines and should each refer to a different real-world entity. For example, there will be only one entry for a specific victim, no matter how many different ways he or she is referred to in the story. On the other hand, if three different people are killed in a bombing, then all three victims should be listed in the template, on separate lines. If the story contains no information corresponding to a slot (e.g., no victims are mentioned), then the slot should be filled with a hyphen (-).

The correct output for the DEV-MUC3-0644 story is:

ID:	DEV-MUC3-0644
INCIDENT:	BOMBING
WEAPON:	BOMB
PERP INDIV:	GUERRILLAS
PERP ORG:	MAOIST POPULAR LIBERATION ARMY
TARGET:	PEASANT HOME
VICTIM:	PEOPLE

This template format is a shortened form of the templates used in the Third and Fourth Message Understanding Conferences (MUC-3 and MUC-4). The organizers provided a data set consisting of 1700 news articles and answer templates. We will be using a subset of this data containing stories that describe exactly one terrorist event.

The answer templates will be formatted exactly the same way that your system output should be formatted, with one exception. The answer templates may contain several alternative answers that are equally acceptable. In this case, the alternative answers will be separated by a slash ("/"). This often happens in cases of coreference, when a victim is referred to in multiple ways. Your system will be given credit for a correct answer if it finds any of the acceptable alternatives listed in the answer template. **However, your system should not produce alternative answers! There should be no slashes in your output templates. Each line in your output template should be a completely separate entry.**

Not all anaphora are listed in the answer templates, though. For example, pronouns are virtually never legitimate answers by themselves because they are not well-defined outside the context of the story. (That is, it's not a satisfying answer to say that "it" was used as a weapon!) But alternative acceptable answers are possible when multiple descriptive terms are mentioned in the story (e.g., "Fred Smith" and "President Smith").

Detailed Example of Answer (Gold) and Output (System) Templates

As an example, consider story DEV-MUC3-0046:

> DEV-MUC3-0046 (NOSC)
>
> TEGUCIGALPA, 26 JAN 89 – [COMMUNIQUE] [ANTICOMMUNIST ACTION ALLIANCE (AAA)] [TEXT] THE ANTICOMMUNIST ACTION ALLIANCE [ALIANZA DE ACCION ANTICOMUNISTA] REPORTS THE FOLLOWING TO THE HONDURAN PEOPLE, IN PARTICULAR, AND TO THE INTERNATIONAL COMMUNITY IN GENERAL:
>
> 1. DURING A MEETING TODAY, THE DEMOCRATIC PATRIOTIC COMMITTEES AND OUR ORGANIZATION'S DISCIPLINARY TRIBUNAL TRIED AND SENTENCED TO DEATH THE FOLLOWING TERRORISTS OF INTERNATIONAL COMMUNISM: JORGE ARTURO REINA, JUAN ALMENDAREZ BONILLA, RAMON CUSTODIO LOPEZ, ANIBAL PUERTO, AND HECTOR HERNANDEZ.
>
> 2. THESE PEOPLE WILL BE EXECUTED BEGINNING ON THIS DATE FOR THE SAKE OF LOVE AND FREEDOM. DEATH TO COMMUNISM!

The gold answer template for DEV-MUC3-0046 is shown below. The event type is ATTACK because this story describes a murder. (Some of the stories, like this one, are terrorist communiques, which means that the terrorist group wrote the article announcing its actions.) There are two perpetrators that should be extracted for the PERP INDIV slot: "DEMOCRATIC PATRIOTIC COMMITTEES" and "OUR ORGANIZATION'S DISCIPLINARY TRIBUNAL", and there are five victims that need to be extracted. In the PERP ORG slot, there is only one organization that needs to be extracted but it is referred to in three different ways. Any of the three strings ("ANTICOMMUNIST ACTION ALLIANCE", "AAA", "ALIANZA DE ACCION ANTICOMUNISTA") is an acceptable answer.

ID:	DEV-MUC3-0046
INCIDENT:	ATTACK
WEAPON:	-
PERP INDIV:	DEMOCRATIC PATRIOTIC COMMITTEES
	OUR ORGANIZATION'S DISCIPLINARY TRIBUNAL
PERP ORG:	ANTICOMMUNIST ACTION ALLIANCE / AAA / ALIANZA DE ACCION ANTICOMUNISTA
TARGET:	-
VICTIM:	JORGE ARTURO REINA
	JUAN ALMENDAREZ BONILLA
	RAMON CUSTODIO LOPEZ
	ANIBAL PUERTO
	HECTOR HERNANDEZ

Hypothetical system output for DEV-MUC3-0046 might look like this:

ID:	DEV-MUC3-0046
INCIDENT:	ATTACK
WEAPON:	-
PERP INDIV:	DEMOCRATIC PATRIOTIC COMMITTEES
PERP ORG:	ALIANZA DE ACCION ANTICOMUNISTA
TARGET:	-
VICTIM:	JORGE ARTURO REINA
	JUAN ALMENDAREZ BONILLA
	RAMON CUSTODIO LOPEZ
	ANIBAL PUERTO
	HECTOR HERNANDEZ
	THESE PEOPLE

This (hypothetical) system correctly identified the incident type, found one of the two individual perpetrators, found the perpetrator organization, and identified all of the 5 victims. However, it failed to extract one of the individual perpetrators ("OUR ORGANIZATION'S DISCIPLINARY TRIBUNAL") and extracted one spurious answer ("THESE PEOPLE").

Input

Your IE system should accept a single input file as a command-line argument, which will contain the texts to be processed. We should be able to run your program like this:

infoextract <textfile>

The input file will contain multiple news stories that will be appended together. Each story will begin with one of four possible headers: DEV-MUC3-XXXX, TST1-MUC3-XXXX, or TST2-MUC4-XXXX (where the X's are digits), so you can use the headers to determine where one story ends and another begins. A sample input file is available on CANVAS.

Output

As output, your IE system should produce a set of answer templates, one per story. Your system should always produce one answer template for a story and the ID slot should always be filled.

Your system should print the answer templates to a single file that has the same name as the input file but with an added extension of ".templates". For example, if the input file is called "news.txt" then the output file should be named "news.txt.templates". **The answer templates should appear in the output file in exactly the same order as the texts in the input file.** That is, if the input file contains three news stories with the id numbers DEV-MUC3-0001, DEV-MUC3-0002, DEV-MUC3-0003 (in that order) then the templates in the output file should appear in the order DEV-MUC3-0001, DEV-MUC3-0002, DEV-MUC3-0003.

The Data Sets

You will be given three sets of data at different points in the project.

Development Set: 329 stories and gold answer templates

Test Set #1: 100 stories and gold answer templates

Test Set #2: 100 stories and gold answer templates

Project Phases

The project will involve three phases:

Development Phase: The **Development Set** is available on our CANVAS site, which you can use to design and evaluate your IE system. You may use these stories and answer templates in any way that you wish. In addition, the *scoring program* that we will use to evaluate your IE systems is available on CANVAS. You can use this scoring program to assess the performance of your system as you experiment with different ideas.

Midpoint Evaluation: On November 7, there will be a midpoint evaluation of all the IE systems. Each team must submit their code and we will evaluate each IE system on a brand new data set, called **Test Set #1**. Once the midpoint evaluation is over, we will release Test Set #1 and you can try to improve the performance of your system on those articles or use them as additional training data.

Final Evaluation: On November 28, each team will submit the final code for their IE system. We will run your IE system on a different brand new data set called **Test Set #2**.

External Software & Data

You may use external software packages and data resources for your project, as long as the following criteria are met:

- You may NOT use any external software that performs event extraction! If we discover that your submitted system uses any external system or code that performs event extraction, you will get a zero for the project.
- You must fully acknowledge in your final presentations/posters all of the external software and data resources that you used in your system.
- We must be able to run all of your software (your own and external resources) on the linux-based CADE machines. This means either including the software in your code submission, or installing it in your own CADE directory and giving us full permission to access it from there. Feel free to discuss options with the TAs.
- You MAY use external NLP software that performs basic NLP functionality, including tokenizers, sentence splitters, part-of-speech taggers, syntactic parsers, coreference resolvers, and general-purpose semantic dictionaries such as WordNet.

If you are uncertain about whether a specific resource is acceptable to use, please ask us!

Machine Learning

You do not need to use machine learning (ML) for this project. But you are welcome to do so if you wish. If you choose to use ML:

- You MAY use external ML software packages.
- You may NOT use any ML models that have been trained for event extraction. If you create an ML model for event extraction, you must train it yourself.
- You MAY use the Development Set as training data for both the Midpoint and Final Evaluations. You may also use Test Set #1 as additional training data for the Final Evaluation, if you wish.
- You may NOT use any additional sources of training data. If you submit an ML model that has been trained on external data, your IE system will be disqualified and you will get a zero for the project. The reason is that we want to level the playing field for all teams in the class, so that everyone is using exactly the same data.

Evaluation

The performance of each IE system will be evaluated using the **F-measure** statistic, which combines *recall* and *precision* in a single metric.

Recall (R): the number of correct items extracted by your system divided by the number of items in the answer template. An extracted string is correct only if it exactly matches one of the strings in the answer template.

Precision (P): the number of correct items extracted by your system divided by the total number of items extracted by your system.

F-measure: $\mathbf{F(R, P)} = \frac{2 \times P \times R}{P + R}$
This formula tries to find a good balance between recall and precision. (It is the harmonic mean of recall and precision.) *The final performance of each system will be judged based on its F-measure score.*

The scoring program that we will use to evaluate your IE systems is available on CANVAS, so you will know exactly how we will be computing the scores. We encourage you to use it during system development as well, to track the performance of your system as you make changes and try to improve it.

Schedule

The schedule for the projects is shown below:

October 18 by 11:00pm: Team member information is due.

November 7 by 11:00pm: Midpoint evaluation on Test Set #1.

November 28 by 12:00pm (noon): Final evaluation on Test Set #2.

December 4: In-class project presentations (top 5 teams)

December 6: In-class poster presentations (remaining teams)

Details on the project presentations will be forthcoming.

Grading

Each project will be graded according to the following criteria:

- 30% of the grade will be based on the performance of your IE system on Test Set #1 during the midpoint evaluation.
- 65% of the grade will be based on the performance of your IE system on Test Set #2 during the final evaluation.
- 5% of the grade will be based on your project presentation, which will be an in-class presentation for the 5 top-performing teams and a poster presentation for remaining teams. All teams will be required to submit the slides used in their presentations for grading.

To determine the grades for the midpoint and final evaluations, each team will be ranked based on the performance of their system relative to the other IE systems. The teams will then be clustered (manually) so that teams whose IE systems produced similar scores will get similar grades.

Note that it is fine to share ideas with other teams (but not code!), and to compare your system's performance with other teams. But if your team is doing almost exactly the same thing as many other teams, chances are your system will end up in the middle of the rankings. To distinguish yourself and stand apart in the rankings, we encourage teams to try different things!

IMPORTANT: For each aspect of the project, if both team members contribute (roughly) equally, as I expect in most cases, then both people will get the same grade. But if one teammate contributes very little, and substantially less than the other teammate, then that person will get a lower grade than the person who put in substantially more time and effort.

CAVEAT AND ENCOURAGEMENT

Building an effective information extraction system is hard! I do not expect that most systems will get high recall and precision (although I am hoping that some systems will do reasonably well!). Your goal is to build the best IE system that you can. The final grading will be based on how well your system does relative to the other teams' systems, but it is not the case that the highest ranked system will get an automatic 'A' or that the lowest ranked system will get an automatic 'E'. If every team produces a system that works wonderfully well, then I will be thrilled to give everyone an 'A' on their project! If, at the other extreme, no team generates a system that works at all, then I would be forced to give every team a failing grade. I hope that this project will be fun, and it will definitely give you hands-on experience building a real NLP application. I hope that the "competitive spirit" will energize everyone to work hard and produce interesting IE systems!