Charles Drews (csd305@nyu.edu, N11539474)

Term Project

Natural Language Processing, CSCI.GA.2590, Spring 2014

Project Goals

The intent of my project is to perform information extraction on a relatively narrow sublanguage: press releases announcing new albums from musical artists. The information to be extracted is as follows:

- Name of the group or solo artist releasing an album
- Name of the album being announced
- Release date of the album, if specified
- Record company releasing the album, if specified

My goal is to process articles in Jet, and have the data points listed above be automatically found and printed to standard output. If one or more of the pieces of information are not included in a particular article (e.g. if it is an independently released album and no record company is mentioned) then that data item will not be printed.

Attachments

In addition to this report, please find attached the following files:

- `album.jet`
- `artist_names.dict`
- `record_co_names.dict`
- `albumConcepts.hrc`
- `albumChunkPatternsPA.txt`
- `albumPatterns.txt`
- `announcements1.txt`
- `announcements2.txt`

Input Data

My input data consists of articles found via Google searches and pasted into text files: `announcements1.txt` contains a selection of 25 training articles, and `announcements2.txt` contains a selection of 15 testing articles. The former was used in the formation of my patterns (described below under Methodology) and the latter was used to test the generality of those patterns (described below under Results).

My Google search phrases were "album press release" and "new album press release." I found that searching for "new album announcement" yielded mostly posts on fan blogs, which tended to be written in a very informal style with very unpredictable patterns. Searching specifically for press releases proved helpful, as

press releases tend to be more formal and predictable. This means my project is focused specifically on press releases, and not on other forms of album announcement news like blog posts.

Also, some of the press releases I encountered had an unfortunate habit of putting the album announcement information in the title of the article, and not restating it in the body of the article. I skipped such articles (fortunately only a minority of my search results) for the following reason - typically every word in an article headline is capitalized, which means there is no way to distinguish between a possible album name and any other word in the title. Therefore, I chose to analyze only the body of each article and not the title (consistent with the approach used in homework assignment #7 for analyzing appointment events). This meant it was necessary to exclude press releases that stated vital information only in the title and not again in the body.

It is also worth noting that some articles did not contain all four data points listed above. About a third did not list a record company, either because it was describing an independently released album, or because it was a press release originating from the artist or group themselves and simply chose not to mention the record company. Also, some did not mention a release date. This is possibly because the announcement was issued after the release of the album, or because the date was not yet known. Either way, I took into account how many data points were actually present in each article when assessing how accurate my system was at extracting data. In other words, I did not penalize my accuracy score for not finding a record company name if one was not provided by the article.

The only modifications I made to the input text, aside from pasting them from a website into a text file, was to replace non-ASCII characters with their nearest-match ASCII characters. I noticed that curved quotation marks and apostrophes seemed to cause problems, so I replaced them with straight quotation marks and apostrophes. I also noticed that accented vowels such as "é" seemed to cause difficulty, so I replaced them with their non-accented counterparts.

Methodology

This project is about finding names and dates. Professor Grishman's initial warning was that the Jet name tagger may not be helpful with artist and album names which may be in all caps, use strange symbols, or be just single names (as opposed to Firstname Lastname). My initial tests using the name tagger proved this concern to be correct. Professor Grishman suggested assuming that artist and record company names are already known, and go in a dictionary. I followed this suggestion and created two custom dictionary files: `artist_names.dict` and `record_co_names.dict`. My Jet properties file, `album.jet`, calls these extra dictionaries after the standard `Jet4.dict` using the `lexLookup` action.

I used the `tagTimex` action with the `time_rules.yaml` file to tag dates. I set the reference time simply as 2014-05-06 in `album.jet` rather than try to automatically update the reference date based on the publication date of each article. I felt this was reasonable for two reasons: first, some articles did not provide a publication date, and second, normalizing dates based on a reference date is a sufficiently difficult task that it was suggested as an entire project. I didn't think it was necessary to tackle that task on top of my album information extraction goals. Therefore, the album release date output from my system may include relative dates if an article provided only a relative and not absolute date.

After tokenizing, performing dictionary lookups, pruning tags, and tagging dates, then I applied all my patterns. These are divided into two files. The first is named `albumChunkPatternsPA.txt` and contains the patterns that create my ngroup, vgroup, name, and date tags. The actions in my properties file that correspond to this patterns file are `pat(dates)`, `pat(names)`, and `pat(chunks)`.

The PA feature is used for ngroups and vgroups, which allowed me to preserve head word information for each group. The ngroups were given one additional feature, named st, which has the value true if the adjectives before the head noun include "self titled" or "self-titled" and otherwise has the value false. This helped me to later detect announcements of self-titled albums where the words "self titled" were often buried in the middle of an ngroup.

The name tags are applied to artist names, record company names, and possible album names. The latter is defined as sequences of capitalized words, interspersed with uncapitalized filler words ("a", "the", etc.) and

apostrophes (allowing for 'Round in place of Around, for example), which do not appear in either the artist name dictionary or the record company name dictionary.

The date tags are applied to TIMEX2 constituents, with a bit of extra cleanup. I noticed that day numbers appearing before a month name were ignored and not included in the TIMEX2 constituent. For example "5th May" would be "5th <TIMEX2>May</TIMEX2>" which of course makes pattern matching a bit difficult because there is now an unexpected "5th" next to the date constituent. To resolve this issue, I made sure my date tags would include both the TIMEX2 constituent and any day indicators appearing either directly before or after (with possible comma separations). Furthermore, some articles gave very specific dates, such as "Saturday, April 1" which would result in two TIMEX2 constituents, one for "Saturday" and one for "April 1." I made sure to combine these into a single date tag so as to simplify pattern matching later.

My second pattern file is named `albumPatterns.txt`, and is where I specify the patterns which extract the necessary data points. The actions in my properties file that correspond to this patterns file are `pat(artist)`, `pat(album)`, `pat(record-co)`, and `pat(release-date)`. This file relies heavily on my concepts file, which is named `albumConcepts.hrc`. The top section of `albumPatterns.txt` defines and gives simple names to various vgroups and ngroups with heads that correspond to a specific concept listed in `albumConcepts.hrc`. For example, "`vg-announce`" is defined, via the concepts mechanism, to be all vgroups with a head verb in my list of announcement verbs. This greatly simplified the rest of my patterns, so I didn't have to use the "`...[head?isa(c...`" syntax every time I wanted to refer to one of these vgroups or ngroups.

Another important aspect of this pattern file is that I have separate pattern sets for each of the four data items I am seeking to extract. I began the project by using just a single pattern set which tried to find all four data points within one (sometimes quite long) pattern, then print out all four data points at once. This approach quickly became untenable as I realized just how many possible phrasings there are for an album announcement, and because sometimes one of the four data points is located several sentences away from the other points. Therefore, I decided to use separate pattern sets, meaning first I scan for a relatively shorter pattern containing a use of an artist name in the context of an announcement, then print that artist name. Then I move on to scanning for a pattern containing an album name in the context of an announcement, then print that album name, and so on. In this fashion, finding each data point is a separate activity.

This approach has pros and cons. An advantage is that while a pattern long enough to contain all four data points can have almost infinite variation, if I instead consider a smaller pattern fragment containing just one of the data points (or possibly others if they are nearby and help identify the target data point) there are fewer variations to consider. Also, this allows me to find all four data points even in situations where they are not located nearby each other in the article. For example, one article did not provide an album release date anywhere near the artist name or album name information, but several paragraphs later gave a date for the release party, which in my estimation means it is a release date.

The downside to this approach is that sometimes my system extracts the same data point from two different locations in an article, and prints the same data point twice to standard output. Fortunately, this was a relatively rare occurrence, and in my opinion does not inhibit the success of the information extraction, since the correct data point is indeed provided. I took some steps to reduce the frequency of this occurrence as well. Artist and record company names, as described above, appear in dictionaries and are obviously easy to spot by their tags. I could have simply printed an artist name to standard output whenever it appears in the article, but to me that seemed inappropriate; we really just want to isolate the artist name one time in the article, where it appears in the context of the actual album announcement. The same situation applies for record company names. Therefore, `albumPatterns.txt` includes many patterns for locating only the relevant instances of artist and record company names in order to ensure they are reported to standard output just once and not every time they appear.

My patterns for album name look for a possible name, as tagged by `albumChunkPatternsPA.txt,` in a variety of specified contexts that make it clear it represents an album name. My release date patterns perform the same task for dates. All of these pattern sets are called in sequence by `album.jet`. All patterns were developed using my training article set (`announcements1.txt`). I worked my way through the training articles until I had as close to 100% precision as I could achieve, all the while looking for opportunities to make the patterns more generalized and applicable to wider variations in phrasing. Then I applied my

system to my testing article set (`announcements2.txt`) to see if my patterns were general enough to extract information from new inputs.

Results

I achieved 97.8% recall and 88.8% precision, for an F-measure of 93.0% on my training article set (`announcements1.txt`). Hand annotation yielded 89 total tags in my training set. Jet annotation yielded 98 tags, of which 87 were correct and 11 were incorrect. The recall was lower than the precision because my system produced some extra annotations, some of which were repeats of correct annotations, and some of which were totally incorrect. To be fair, I counted both types as incorrect when calculation my precision.

I achieved 73.1% recall and 92.7% precision, for an F-measure of 81.7% on my testing article set (`announcements2.txt`). Hand annotation yielded 52 total tags in my testing set. Jet annotation yielded 41 tags, of which 3 were correct and 38 were incorrect. Unsurprisingly, my recall dropped relative to the training set because the articles of the testing set contained new phrasings I had not previously considered or allowed for in my patterns. On the other hand, my precision was actually better than what I achieved with my training set, which means I am not achieving my recall by simply tagging everything in sight, rather my patterns are actually working fairly precisely. Overall, I am pleased with my system's performance with the testing articles, and think an F-measure of 81.7% is a success.

Below is a listing of the all data points found in each article of my training and testing sets by my hand annotation, and by my Jet system, with commentary:

Training Article	Hand Annotation	Jet Annotation	Comments
1	Artist: HAMMERFALL Album: Infected Release Date: May, 20th	Artist: HAMMERFALL Album: Infected Release Date: May, 20th	
2	Artist: Michael Jackson Album: XSCAPE Record Company: Epic Records Release Date: May 13th	Artist: Michael Jackson Album: XSCAPE Record Company: Epic Records Release Date: May 13th	
3	Artist: Garbage Album: Not Your Kind Of People Record Company: STUNVOLUME Release Date: May 15, 2012	Artist: Garbage Album: Not Your Kind Of People Record Company: STUNVOLUME Release Date: May 15, 2012	
4	Artist: Scott Stapp Album: Proof of Life Record Company: Wind-up Records Release Date: November 5, 2013	Artist: Scott Stapp Album: Proof of Life Record Company: Wind-up Records Release Date: November 5, 2013	
5	Artist: Atash Album: Everything Is Music Release Date: March 25, 2014	Artist: Atash Album: Everything Is Music Release Date: March 25, 2014	
6	Artist: Roger Taylor Album: Fun On Earth Release Date: October	Artist: Roger Taylor Album: Fun On Earth Release Date: October	

Training Article	Hand Annotation	Jet Annotation	Comments
7	Artist: Greg Laswell Album: Landline Record Company: Vanguard Records Release Date: April 24th	Artist: Greg Laswell Album: Landline Album: Landline Record Company: Vanguard Records Release Date: April 24th	My system printed the album name twice. It is possible that one pattern matched in mention in the article and a different pattern matched a separate mention.
8	Artist: Bob Dylan Album: Tempest Record Company: Columbia Records Release Date: September 11, 2012	Artist: Bob Dylan Album: Tempest Record Company: Columbia Records Release Date: September 11, 2012	
9	Artist: Nathan Jess Album: Love Stands Forever Record Company: Integrity Music	Artist: Nathan Jess Album: Love Stands Forever Record Company: Integrity Music	
10	Artist: Rye Album: Cumberland Island Release Date: May 14, 2013	Artist: Rye Album: Cumberland Island Release Date: May 14, 2013	
11	Artist: Cilantro Boombox Album: self-titled Release Date: September 14th	Artist: Cilantro Boombox (Album not captured) Release Date: September 14th	My system did not capture the fact that the album is self-titled. Oddly, when I paste the relevant sentence from the article into Jet's input box, it does print "Album: self-titled" but not when processing from the text file.
12	Artist: The Roots Album: undun Record Company: Island Def Jam Release Date: December 6th	Artist: The Roots (Album not captured) Record Company: Island Def Jam Release Date: December 6th	My system failed to capture the album name. This occurred because the name begins with a lowercase letter, and a fundamental assumption of my system is that album names will be capitalized.
13	Artist: Alan Cave Album: TIMELESS Record Company: AC Records Release Date: Saturday, April 19, 2014	Artist: Alan Cave Album: TIMELESS Record Company: AC Records Release Date: Saturday, April 19, 2014	

Training Article	Hand Annotation	Jet Annotation	Comments
14	Artist: Jose James Album: While You Were Sleeping Record Company: Blue Note Records Release Date: June 10	Artist: Jose James Album: While You Were Sleeping Record Company: Blue Note Records Record Company: Blue Note Records Release Date: June 10	My system printed the record company twice. It is possible that one pattern matched in mention in the article and a different pattern matched a separate mention.
15	Artist: Michael W. Smith Album: Sovereign Record Company: Sparrow Records and Capitol Christian Music Group Release Date: May 13	Artist: Michael W. Smith Album: Sovereign Record Company: Sparrow Records and Capitol Christian Music Group Record Company: Sparrow Records and Capitol Christian Music Group Release Date: May 13	My system printed the record company twice. It is possible that one pattern matched in mention in the article and a different pattern matched a separate mention.
16	Artist: Echo & the Bunnymen Album: Meteorites Record Company: 429 Records Release Date: April 28th	Artist: Echo & the Bunnymen Album: Meteorites Record Company: 429 Records Release Date: April 28th Record Company: 429 Records	My system printed the record company twice. It is possible that one pattern matched in mention in the article and a different pattern matched a separate mention.
17	Artist: Avril Lavigne Album: self-titled Record Company: Columbia Records Release Date: September 24th	Artist: Avril Lavigne Album: self-titled Record Company: Columbia Records Release Date: September 24th	
18	Artist: LAIBACH Album: SPECTRE Release Date: 3 March 2014	Artist: LAIBACH Album: SPECTRE Release Date: 3 March 2014	
19	Artist: Il Volo Album: We Are Love Record Company: Interscope Records Release Date: November l9th	Artist: Il Volo Album: We Are Love Record Company: Interscope Records Release Date: November l9th	

Training Article	Hand Annotation	Jet Annotation	Comments
20	Artist: Pink Floyd Album: The Dark Side Of The Moon Record Company: EMI Records Release Date: 24th March 2003	Artist: Pink Floyd Album: The Dark Side Of The Moon Record Company: EMI Records Release Date: 24th March 2003 Release Date: 24th March, 1973 Release Date: Friday, April 22	My system printed two extra, erroneous dates because they were mentioned in a context that matched one of my patterns.
21	Artist: Atomic Skunk Album: Alchemy Release Date: April 22, 2011	Artist: Atomic Skunk Album: Alchemy Release Date: April 22, 2011 Album: Alchemy	My system printed the album name twice. It is possible that one pattern matched in mention in the article and a different pattern matched a separate mention.
22	Artist: BLACK STONE CHERRY Album: Magic Mountain Record Company: Roadrunner Records Release Date: 5th May	Artist: BLACK STONE CHERRY Album: Magic Mountain Record Company: Roadrunner Records Release Date: 5th May	
23	Artist: LEATHERWOLF Album: Unchained Live Release Date: December 15	Artist: LEATHERWOLF Album: Unchained Live Release Date: December 15	
24	Artist: Beck Album: Morning Phase Record Company: Capitol Records Release Date: February 2014	Artist: Beck Album: Morning Phase Record Company: Capitol Records Release Date: February 2014 Artist: Beck Release Date: September Artist: Beck Release Date: 2014	My system printed the artist name two extra times. It is possible that one pattern matched in mention in the article and a different pattern matched a separate mention. My system also printed two extra, incorrect dates because they were mentioned in a context that matched one of my patterns.
25	Artist: Jason Lee Greenberg Album: Orisonata	Artist: Jason Lee Greenberg Album: Orisonata	

Below is a listing of the data points found in each article of my testing set by my hand annotation, and by my Jet system, with commentary:

Testing Article	Hand Annotation	Jet Annotation	Comments

Testing Article	Hand Annotation	Jet Annotation	Comments
1	Artist: Gold Panda Album: Half Of Where You Live Record Company: Ghostly International	Artist: Gold Panda Album: Half Of Where You Live Album: Half Of Where You Live Record Company: Ghostly International	My system printed the album name twice. It is possible that one pattern matched in mention in the article and a different pattern matched a separate mention.
2	Artist: Say Anything Album: Hebrews Record Company: Equal Vision Records Release Date: June 10th, 2014	Artist: Say Anything Album: Hebrews Record Company: Equal Vision Records Release Date: June 10th, 2014	
3	Artist: Simone Moreno Album: Planetas Record Company: Soul Dog Records	Artist: Simone Moreno Album: Planetas (Record Company not captured)	None of my patterns matched the phrasing for the release date in this article.
4	Artist: Jay-Z Album: Magna Carta Holy Grail Release Date: July 4	Artist: Jay-Z Album: Magna Carta Holy Grail Release Date: July 4	
5	Artist: The Blessed Broke Album: Ladders Out of Purgatory Release Date: April 22nd	(Artist not captured) (Album not captured) (Release Date not captured)	None of my patterns matched the phrasing for any of the data points in this article.
6	Artist: BelO Album: Natif Natal Record Company: Haitian International Release Date: April 9, 2014	Artist: BelO Album: Natif Natal Record Company: Haitian International (Release Date not captured)	None of my patterns matched the phrasing for the release date in this article.
7	Artist: Sonny Knight & The Lakers Album: I'm Still Here Record Company: Secret Stash Records Release Date: April 29th, 2014	(Artist not captured) Album: I'm Still Here Record Company: Secret Stash Records Release Date: May 3rd	None of my patterns matched the phrasing of the artist name or release date in this article; furthermore, one of my release date patterns matched an incorrect date from the article which should have been ignored.
8	Artist: Joe Crookston Album: Georgia I'm Here Release Date: Saturday	Artist: Joe Crookston Album: Georgia I'm Here Saturday (Release Date not captured)	None of my patterns matched the phrasing of the release date in this article.
9	Artist: Mastodon Album: Once More 'Round The Sun Record Company: Reprise Records Release Date: June 24th	Artist: Mastodon Album: Once More 'Round The Sun Record Company: Reprise Records Release Date: June 24th	

Testing Article	Hand Annotation	Jet Annotation	Comments
10	Artist: Origin Album: Omnipresent Record Company: Nuclear Blast Records Release Date: July 8, 2014	Artist: Origin Album: Omnipresent Record Company: Nuclear Blast Records Release Date: July 8, 2014	
11	Artist: Porter Robinson Album: Worlds Release Date: Aug. 12	Artist: Porter Robinson Album: Worlds Release Date: Aug. 12	
12	Artist: Seth Boyer Album: Half Lonely Release Date: Friday, April 25	Artist: Seth Boyer Album: Half Lonely Release Date: Friday, April 25	
13	Artist: Ahmad Jamal Album: Saturday Morning Record Company: Jazzbook records Release Date: September 2013	Artist: Ahmad Jamal (Album not capture) (Record Company not captured) Release Date: September 2013	None of my patterns matched the phrasing of the album name and record company name in this article.
14	Artist: Nick Cave & the Bad Seeds Album: PUSH THE SKY AWAY Release Date: 18 February 2013	(Artist not captured) (Album not captured) (Release Date not captured)	The key sentence of this article used the un-tensed "release" which my system did not tag as a verb group. If it had been "releases" or "released" my system would have correctly captured all three data points, but alas that was not the case.
15	Artist: Matisyahu Album: Akeda Record Company: Elm City Music Release Date: June 3	Album: LP Record Company: Elm City Music Release Date: June 3 Artist: Matisyahu	My system failed to capture the correct album name. It also presented the other data points out of order, but they are at least accurate.