Hybrid Information Extraction Systems for Open Data

Class 5: Hybrid IE Systems

Pablo Ariel Duboue, PhD

Curso de Posgrado Cs. de la Computacion
FaMAF-UNC

Why Hybrid IE?

- Use the right tool for the job
  - Simple NEs: REs
  - List-based NEs: dictionaries
    - With suitable filtering
- Pitfall: cascade of errors

Annotations

- This discussion follows "Natural Language Annotation for Machine Learning" by Pustejovsky and Stubbs (2012)
- MATTER Cycle: Model Annotate Train Test Evaluate Revise

Levels of Annotations

- Syntax
- Semantics
- Morphology
- Phonology
- Phonetics
- Lexicon
- Discourse analysis
- Pragmatics
- Text structure analysis

Penn Tree Bank Tagset

1.	CC	Coordinating conjunction	25.	TO	to
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential there	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	Verb, gerund/present participle
6.	IN	Preposition/subord.	30.	VBN	Verb, past participle
218z		conjunction			
7.	JJ	Adjective	31.	VBP	Verb, non-3rd ps. sing. present
8.	JJR	Adjective, comparative	32.	VBZ	Verb, 3rd ps. sing. present
9.	JJS	Adjective, superlative	33.	WDT	wh-determiner
10.	LS	List item marker	34.	WP	wh-pronoun
11.	MD	Modal	35.	WP	Possessive wh-pronoun
12.	NN	Noun, singular or mass	36.	WRB	wh-adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence-final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semi-colon
18.	PRP	Personal pronoun	42.	(	Left bracket character
19.	PP	Possessive pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	`	Left open single quote
22.	RBS	Adverb, superlative	46.	"	Left open double quote
23.	RP	Particle	47.	'	Right close single quote
24.	SYM	Symbol	48.	"	Right close double quote

From Pustejovsky and Stubbs (2012)

Model

- Model: a characterization that is more abstract than what is being modeled.
  - The tagset of spec
- With a model, annotation guidelines can be written

Guidelines

- Consuming vs non-consuming tags
- Interannotator agreement, span size.
- Kappa statistic.
- Good agreement means good instructions not good annotations!
- Adjudication for gold standard

Interannotator Agreement

- *"Inter-Coder Agreement for Computational Linguistics"* by Artstein & Poesio (2008)
  - http://aclweb.org/anthology/J/J08/J08-4004.pdf
- *"Assessing Agreement on Classification Tasks: The Kappa Statistic"* by Carletta (1996)
  - http://aclweb.org/anthology-new/J/J96/J96-2004.pdf

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the observed agreement and $P(E)$ is the expected chance agreement for annotators choosing each category the same number of times as they originally did, but choosing each item randomly.

Cohen's Kappa Example

- From http://en.wikipedia.org/wiki/Cohen%27s_kappa

	B yes	B no
A yes	20	5
A no	10	15

- $P(A) = {}^{(20+15)}/_{50} = 0.7$
- $P(E) = P(E, yes) + P(E, no)$, A says yes 50% and B says yes 60%, therefore $P(E, yes) = 0.5 \times 0.6 = 0.3$ and $P(E, no) = 0.5 \times 0.4 = 0.2$ and thus $P(E) = 0.3 + 0.2 = 0.5$

$$\kappa = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

More Than Two Annotators

- Move from a contingency table to an agreement table
  - For each annotated item, how many categories it got (from any annotator)
- The generalization of $P(E)$ for multiple annotators involves computing pairwise agreements.
  - The pairwise average of 2-annotator $P(E)$, to be more precise.
  - See the extended version of Artstein & Poesio (2008) for the exact formula
  - For a python implementation:

`http://nltk.org/_modules/nltk/metrics/agreement.html#AnnotationTask.multi_kappa`

Understanding Kappa Values

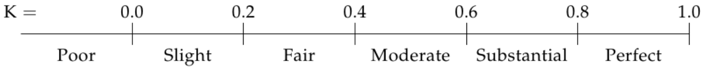


Figure 1
Kappa values and strength of agreement according to Landis and Koch (1977).

(from Artstein & Poesio, 2008)

Statement of Purpose

- Write a statement of purpose (one sentence or so)
  - then expand it with "hows"
- The basic statement should answer:
  - What's the intended use of the annotation
  - What's its overall outcome
  - Where the documents come from
  - What level is being annotated

Refinements

- Informativity (useful annotations) vs correctness (things that can be annotated)
- /TO

Scope of the Annotation Task

- Granularity of tags
- Single sentence vs. multisentence vs multidocument

Scope of the Corpus

- Written material vs transcripts
- Professional prose vs amateur
- Background research:
  - LDC, ELRA, conferences, challenges
- Assembling the dataset: choosing it, getting permissions for re-distribution. Eliciting data.

Annotating

- Deciding what to tell the annotators about metadata (avoiding bias)
- Pre-processing vs clean slate: asking to correct errors plus annotating tend to be overlooked.
  - Better to do in two separate tasks.
- How much to annotate?
  - Sample the corpus so all phenomena are represented
- How the annotators will do their work. What annotation will look like? Different tasks require different representations? annotation exchange format / annotation environment vs machine learning need
  - If the annotation set-up is error prone, errors will creep on top of the annotation errors.
- Inline vs. token standoffs vs. character standoffs annotations.

Full System

- SimpleFrenchSentenceAndToken (java.text.BreakIterator)
- AmountAnnotator (concept)
- NeqConceptAnnotator (ConceptMapper)
- CompanyAnnotator (OpenNLP)
- Combination of companies (custom)
- Reason (RuTA)
- ReasonAnnotator (CRF/ClearTk/Mallet)

Changes to annotate a different NE

- Add type
- Annotate
- Train model
- Set up descriptor

Full Cycle

- Step 1:
  - Baseline system (RuTA + Amount RE + NEQ) on first 36 documents
  - Generate CASes on 36
  - Annotate on 36 / evaluate on 36
- Step 2:
  - Train Company OpenNLP on 36
  - Train Reason CRF on 36
  - Filter NEQ dictionary of spurious matches
  - Generate CASes on 32
  - Annotate on 32 / evaluate on 32

Full Cycle (cont.)

- Step 3:
  - Train Company OpenNLP on 36+32
  - Train Reason CRF on 36+32
  - Filter NEQ dictionary of spurious matches
  - Generate CASes on new 32
  - Annotate on new 32 / evaluate on new 32

Deployment

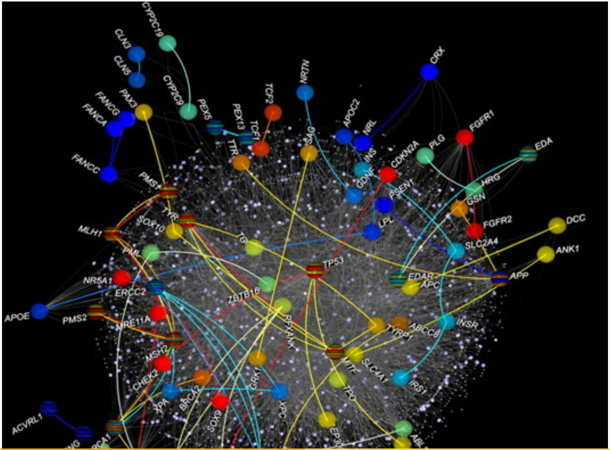
- Robust
- Long term execution
- Memory requirements

Maintenance

- Language drift
- Clustering to maintain NE models

Statistical Inference

- "Network properties of genes harboring inherited disease mutations" by Feldman, Rzhetsky, Vitkup (PNAS 2008)



Natural Language Generation

	A	B	C	D	E	F
1	Item	Category	Price	Profit	Actual Profit	Calories
2	Beer	Beverages	$ 4.00	50%	$ 2.00	200
3	Hamburger	Hot Food	$ 3.00	67%	$ 2.00	320
4	Popcorn	Hot Food	$ 5.00	80%	$ 4.00	500
5	Pizza	Hot Food	$ 2.00	25%	$ 0.50	480
6	Bottled Water	Beverages	$ 3.00	83%	$ 2.50	0
7	Hot Dog	Hot Food	$ 1.50	67%	$ 1.00	265
8	Chocolate Dipped Cone	Frozen Treats	$ 3.00	50%	$ 1.50	300
9	Soda	Beverages	$ 2.50	80%	$ 2.00	120
10	Chocolate Bar	Candy	$ 2.00	75%	$ 1.50	255
11	Hamburger	Hot Food	$ 3.00	67%	$ 2.00	320
12	Beer	Beverages	$ 4.00	50%	$ 2.00	200
13	Hot Dog	Hot Food	$ 1.50	67%	$ 1.00	265
14	Licorice Rope	Candy	$ 2.00	50%	$ 1.00	280
15	Chocolate Dipped Cone	Frozen Treats	$ 3.00	50%	$ 1.50	300
16	Nachos	Hot Food	$ 3.00	50%	$ 1.50	560
17	Pizza	Hot Food	$ 2.00	25%	$ 0.50	480
18	Beer	Beverages	$ 4.00	50%	$ 2.00	200

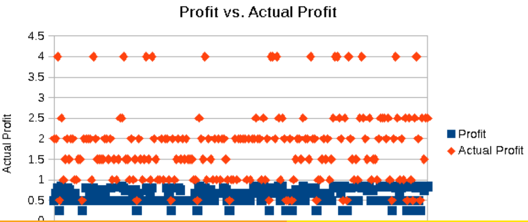
Natural Language Generation (cont.)

Automated Data Analysis of Concessions.xls

The spreadsheet has 199 rows and six columns including a header row.

Column A ('Item ') is non-numeric, with string type. Column B ('Category ') is non-numeric, with type string. Column F ('Calories ') is non-numeric, with type formula.

Column C ('Price ') has a mean of 2.83, a median of 3 and standard deviation of 3. It is very skewed with a strong variance. It is highly correlated with Column E ('Actual Profit').
Column D ('Profit ') has a mean of 0.62, a median of 0.67 and standard deviation of 0.67. It is not skewed at all with a small variance. The correlation with Column E ('Actual Profit') is in the figure.



Other Applications

- PLN FaMAF work with documents from dirty war


LaVoz NOTICIAS MUNDO D VOS ESTILO CLASIFICADOS SERVICIOS MÁS ED.DIGITAL

CIUDADANOS 25/08/2014 00:02

Software cordobés para procesar documentos de la dictadura

Investigadores de Famaf elaboran novedosas herramientas informáticas que facilitan el trabajo del Archivo Provincial de la Memoria.

Científicos comprometidos. Paula Estrella, Martín Domínguez y Franco Luque, investigadores de Famaf (La Voz/Facundo Luque).

ACERCA DEL **AUTOR**

Javier Cámara
Periodista de Política y Negocios

TEMAS DEL **DÍA**

Research topics

- Multilingual
- Multi-document
- Open Domain
- Deep Learning (character-based)
- Unsupervised

People and Groups

- Ralph Grishman and NYU
- Heng Ji and RPI
- Oren Etzioni and UWash
- Ralph Weischedel and BBN

Heng Ji Presentation

- "Information Extraction: Techniques, Advances and Challenges ". Invited Lecture at NAACL Summer School, June 2012.

Open IE

- *"Open information extraction to KBP relations in 3 hours"* by Soderland et al. (Text Analysis Conference. 2013)
- Customizing a general, Open IE System to KBP in 12hs writing mapping **rules**
- Extractor
  - 3hs prec 0.79
  - 12 prec 0.80
- Great example of a hybrid system

Open IE (cont.)

- Open IE expresses the relations textually so multiple "relations" have to be mapped to the actual, ontological relation

Open IE tuples	KBP relations
(Steve Jobs, **died of**, cancer) (Steve Jobs, **succumbed to**, cancer) (Steve Jobs, **lost his battle** to, cancer)	per:cause_of_death
(Nasrallah, **is leader of**, Hezbollah) (Hezbollah, **headed by**, Nasrallah) (Nasrallah, **is Secretary-General of**, Hezbollah)	org:top_members _employees

From Soderland et al. (2013), Fig. 1

https://github.com/knowitall/openie

Open IE (cont.)

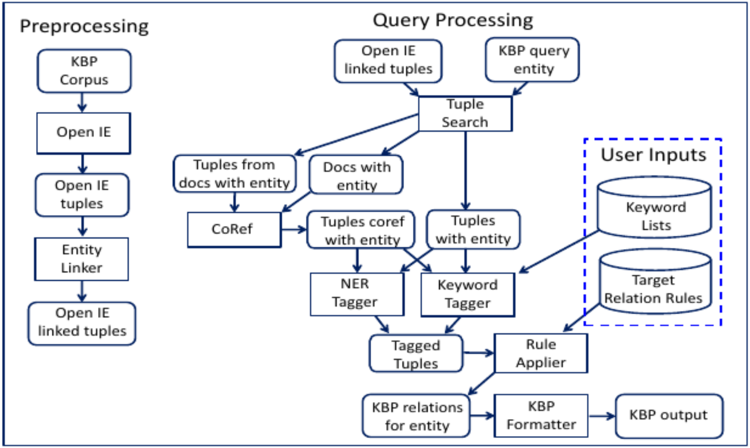
- Zipf says: map the top forms, forget the rest
- Limited training is not enough to learn quality rules, even with active learning (Soderland et al., 2010)
  - More enfasis to rules
- Sample rule:

Terms in Rule	Example
Target relation:	per:employee_or_member_of
Functional?:	No
Query entity in:	Arg1
Slotfill in:	Arg2
Slotfill type:	Organization
Arg1 terms:	-
Relation terms:	appointed
Arg2 terms:	<JobTitle> of

(Smith, was appointed, Acting Director of Acme Corporation)

per:employee_or_member_of (Smith, Acme Corporation)

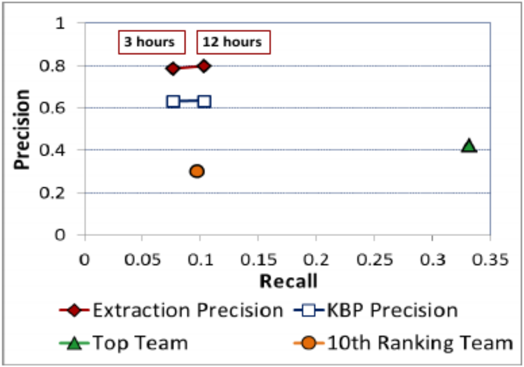
Architecture



From Soderland et al. (2013), Fig. 4

Performance

- 12hs: 16 rules per relation (5x the 3h set)
  - Only 35% increase in recall



From Soderland et al. (2013), Fig. 5

Error Analysis (precision)

- 31% seem correct to them
- 23% rules overgeneralized
- 19% rules matched a non-head term
- 15% errors in the Open IE extractor
- 12% correference errors

Error Analysis (recall)

- Evaluated on a random sample of sentences
- 42% the information was there, a rule was lacking
- 16% the extractor truncated arguments
- 10% Open IE fails to identify a noun-noun relation
- 10% problems due to syntactic complexity
- 22% other

Character-based Deep Learning

- "Deep Learning for Character-based Information Extraction"by Yanjun Qi et al. (ECIR 2014)
- 1.3M Chinese NER

LSTM-CRF

- "Deep Learning for Character-based Information Extraction"by Yanjun Qi et al. (ECIR 2014)
  - 1.3M Chinese NER