Machine Learning for Coreference Resolution

Isabelle Tellier

27/02/2015

Introduction : purpose of this presentation

- introduce coreference resolution : a difficult and important NLP task
- introduce supervised machine learning : a general approach, several algorithms
- see how the approach can be applied to the task

Outline

Outline

Coreference Resolution
Definition of Coreference

- Anaphore : non symmetric relation between an anaphoric expression (pronoun...) and its antecedent
- Coreference : relation shared by various expressions referring to a single entity

« bonjour je suis Madame Nom1 et je souhaiterais parler
à Madame Nom2 s'il vous plaît
- oui c'est moi-même, que puis-je pour vous madame ? »

Coreference Resolution

Hardness of the Task (Example from D. Kayser)

The teacher sent **the pupil** to **the principal**.

- **He** had enough of him.
- **He** was talking to his neighbor.
- **He** wanted to see him.

Coreference Resolution

Hardness of the Task (Examples from Rahman & Ng)

James asked **Robert** for a favor, but **he** refused.
James asked Robert for a favor, but **he** was refused.

Keith fired Blaine but **he** did not regret.
Keith fired **Blaine** although **he** is diligent.

Emma did not pass the ball to **Janie**, although **she** was open.
Emma did not pass the ball to Janie, although **she** should have.

Medvedev will cede the presidency to **Putin** because **he** is more popular.
Medvedev will cede the presidency to Putin because **he** is less popular.

Coreference Resolution

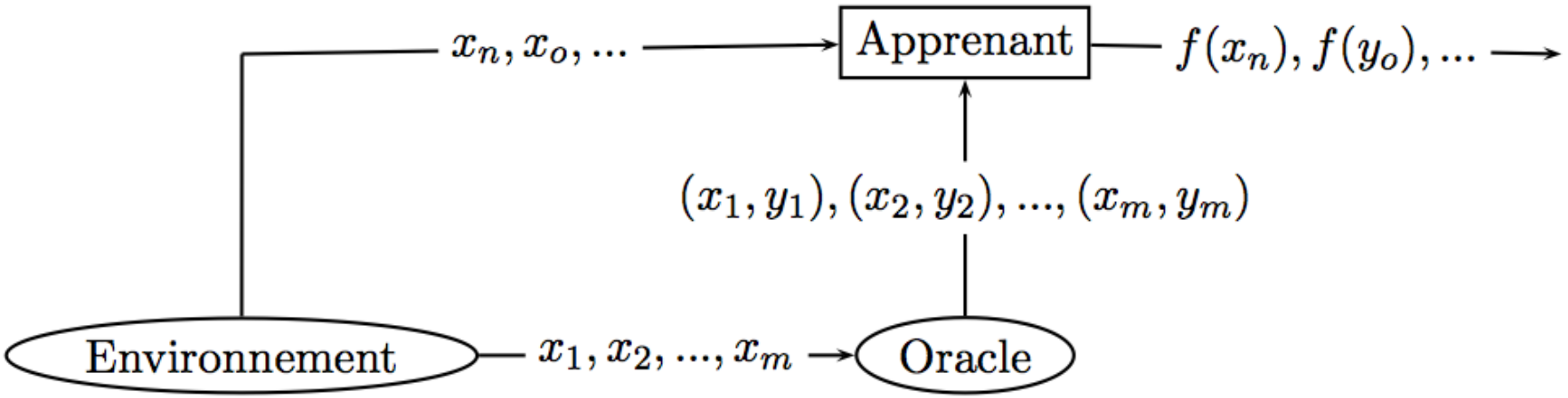
Hardness of the Task, Interest

- well known constraints : agreement in gender and number... are necessary but not enough !
- some examples require large knowledge about the world
- "Winograd Schema Challenge" : possible alternative to the Turing test !
- useful for Information Retrieval (search engines), Information Extraction (automatic filling of a predefined form from a text), Question Answering (answer to a natural language question)
- useful for summary, translation...

Coreference Resolution
Ways of Treating it

- two steps :
  - identifying the entities (and their properties such as gender, number...)
  - grouping co-referential entities together
- various approaches for the second step :
  - rule-based handwritten systems
  - as a clustering problem (unsupervised)
  - by supervised machine learning

Outline



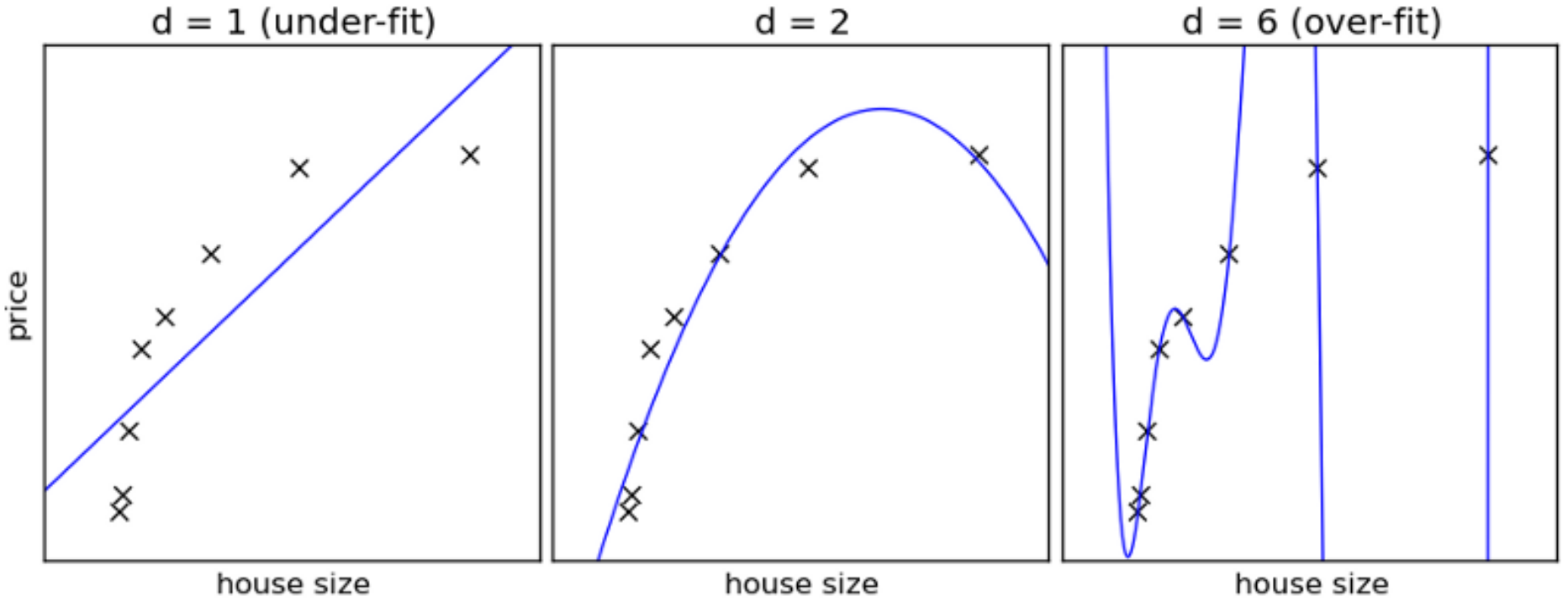
General Schema of Supervised Machine Learning (from Cornuejols & Miclet)

General Properties

Learning (for a machine)

- transforms isolated examples (x, y) into a rule, a function $f(x) = y$
- allows to predict the value of $f(x)$ for some new x
- requires to generalize (learning "by heart" is useless !)... but not too much !
- there are some traps....

General Properties



Three possibles "optimal" functions f for a given set of examples (x, y)

General Properties

Two steps : (1) leaning from examples (2) applying the result to new data.
What is necessary for "learning a function f" from examples (x, y) ?

- (1) a relevant search space of functions
- (1) a distance criterion to optimize
- but to avoid overfitting, the real criterion is (2) ability to predict new values
- thus : need of a learning protocol with a training set (1) distinct from a test set (2)

General Properties

There exist many distinct machine learning algorithms because its choice depends of :

- the nature of x and y (numbers, symbols, strings, vectors...)
- the choice of the search space for f
- the evaluation criterion preferred : ability to predict, computing time, readability/interpretability of the result, incrementally of the result, robustness to new domains...
- No Free Lunch Theorem : no machine learning approach can do better than every other on every possible problem !

ML applied to Classification

- x : data described by features (attributes)
- y : a (symbolic) class chosen among a finite number of possible ones

Applications :

- assurance, bank (ability to repay a loan), medicine (diagnosis, drug effectiveness)...
- text classification : bag of words representation
- text classification according to the domain, the author, the period of writing, the language variant... (any metadata)
- opinion mining !

Demo Weka !

- free, open-source, user-friendly software
- implements many supervised machine learning algorithms
- with various protocols and evaluation metrics
- and many unsupervised (clustering) algorithms
- and many other things !

Three selected algorithms

- Decision Trees (J48) : symbolic approach, readability
- SVM (SMO) : numeric approach, effectiveness (good results), especially for binary classification
- NaiveBayes : statistical approach, efficiency (easy computation)

Outline

ANCOR Corpus

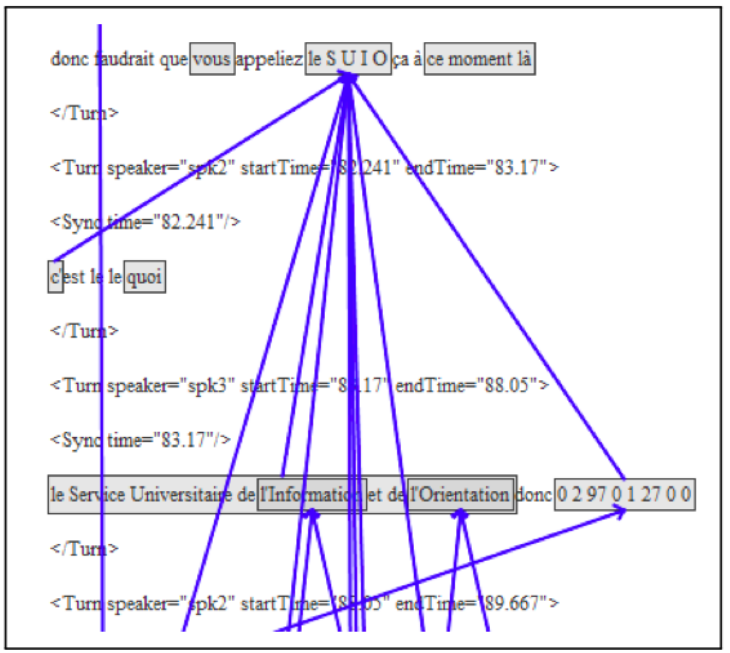
ANCOR : transcribed oral corpora of various origins

Corpus	Discursive Context	Finalisation	Interactivity	Length et Time
ESLO_ANCOR	Interview	Moderate	Weak	417 kMots – 25 h
ESLO_CO2	Interview	Moderate	Weak	35 kMots – 2,5 h
OTG	Oral ialogue	Very strong	Strong	26 kMots – 2 h
Accueil_UBS	Telephonic Dialogue	Quite Strong	Strong	10 kMots – 1 h

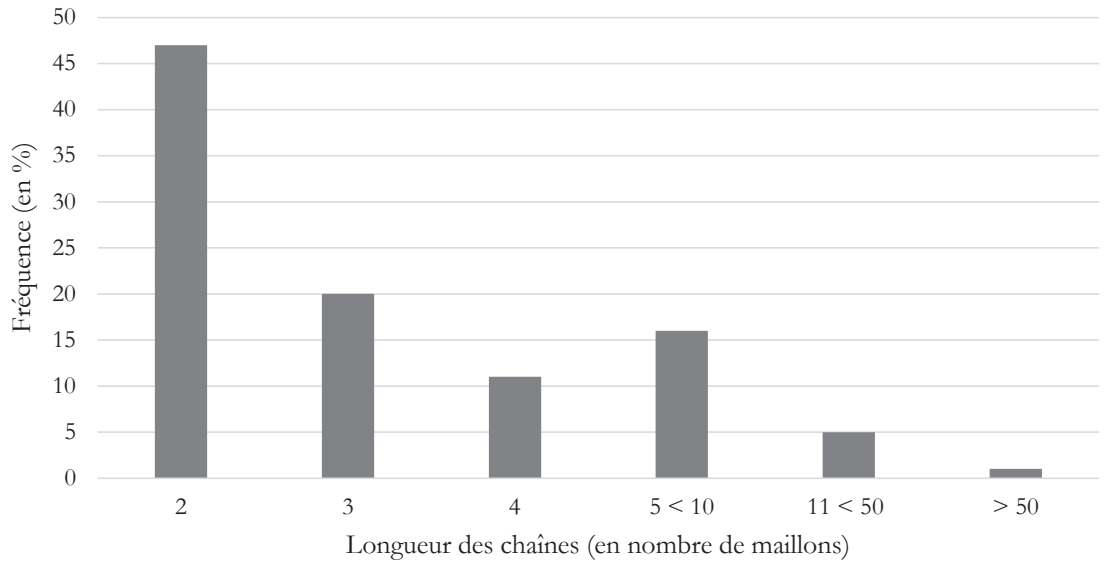
Corpus	ESLO	CO2	OTG	UBS	TOTAL
Entities (of any type)	97 939	8 399	7 462	1 872	115 672
among which new ones (NEW)	*26,8 %*	*32,2 %*	*38,4 %*	*33,7 %*	*28,0 %*
among which belonging to a coreference chain	*73,2 %*	*67,8 %*	*61,6 %*	*66,3 %*	*72,2 %*
Relations (of any type)	44 597	3 670	2 572	655	51 494

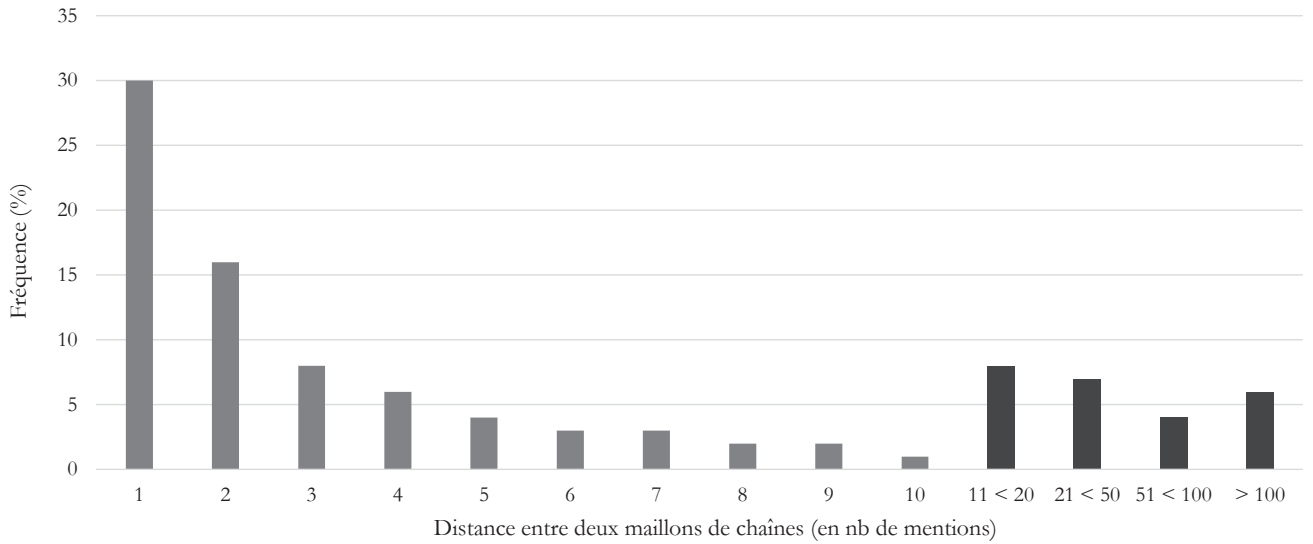
ANCOR Corpus

- manually annotated coreference chains
- each chain link is connected to the first chain link



ANCOR Corpus





Coding Coreference into a Classification Problem

- refering entities e_1, e_2, e_3... are supposed to be recognized in the text, associated with some features (gender, number, lemma...)
- the classification problem is : given a couple (e_i, e_j) of entities, are they coreferent ? (yes/no)
- x : set of features associated with (e_i, e_j)
- y : boolean
- required pre-treatment of ANCOR : each chain link is connected with its previous and/or next chain link (so that the distance/context between the entities of a couple are relevant)

Coding Coreference into a Classification Problem

Possible Features for describing $x = (e_i, e_j)$

- possible non-relational features (related to each entity)
  - gender, number...
  - syntactic category (pronoun, definite NP, demonstrative NP ?)
  - new/old entity ? (not always available !)
  - Named Entity ? Class of NE (person/organization/place/date/numeric) ?
  - semantic category (from Wordnet, ontology...) when available
- possible relational features (related to the couple)
  - equality/inclusion of e_i and e_j as strings, rate of common tokens/substrings
  - for any non-relational feature : equality of their values for e_i and e_j
  - distance in characters, words, sentences, paragraph, speech turns...
  - embedding : [[my neighbor]$_i$'s cat]$_j$ (entities cannot co-refer)
  - equality/inclusion of left/right contexts (in terms of tokens/words)
  - equality of the speaker (in dialogues)

Coding Coreference into a Classification Problem

Reconstructing a coreference chain

- reference data are coreference chains and not couples
- necessity to take into account the fact that coreference is transitive

$\{a_1, b_1, b_2, a_2, c_1, b_3, c_2, a_3\}$

IF $\{a_1, a_2\}$ = COREF AND $\{a_2, a_3\}$ = COREF,
THEN $\{a_1, a_3\}$ = COREF
So, the chain is = $\{a_1, a_2, a_3\}$

Coding Coreference into a Classification Problem

Evaluation Metrics

T_m = reference chain
S_m = predicted chain

$T_m = \{\{1, 2, 4, 7\}, \{3\}, \{5, 6\}\}$

$S_m = \{\{1, 2, 3, 4, 7\}, \{5, 6\}\}$

MUC

$$R = \frac{\#\text{ common links in true and system partition}}{\#\text{ minimum links for true partition}}$$

$$P = \frac{\#\text{ common links in true and system partition}}{\#\text{ minimum links for system partition}}$$

B^3

$$R = \frac{\sum_{i=1}^{n} \frac{\#\text{ common mentions in true and system entity of mention}_i}{\#\text{ mentions in true entity of mention}_i}}{n}$$

$$P = \frac{\sum_{i=1}^{n} \frac{\#\text{ common mentions in true and system entity of mention}_i}{\#\text{ mentions in system entity of mention}_i}}{n}$$

BLANC

$R = \frac{R_c + R_n}{2}$; $P = \frac{P_c + P_n}{2}$

CEAF

complex computation...

Variants of the Problem

Choice of the Set of Features

- all features (30)
- only relational features (18)
- without oral (equality of speaker and speech turn distance) related features (28)

Variants of the Problem

Generating Negative Examples and Training Sets

- problem : negative examples are not "given"
- for any two entities, it is more likely that they are not coreferent

$\{a_1, b_1, b_2, b_3, a_2, c_1, c_2, a_3\}$

$\{a_1,a_2\}$ -> $\{b_1,a_2\}$, $\{b_2,a_2\}$, $\{b_3,a_2\}$
$\{a_2,a_3\}$ -> $\{c_1,a_3\}$, $\{c_2,a_3\}$

- big training set : 142 498 couples (24 620 coref, 117 878 not)
- medium training set : 101 919 couples (17 844 coref, 84 075 not)
- small training set : 71 881 couples (11 908 cored, 59 973 not)

Variants of the Problem

Choice of the learning algorithm

- Decision Trees (J48) : symbolic approach, readability
- SVM (SMO) : numeric approach, effectiveness (good results), especially for binary classification
- NaiveBayes : statistical approach, efficiency (easy computation)

Experiments and Results with ANCOR

		small_trainingSet			*medium_trSet*		*big_trSet*	
		NBayes	SVM	C4.5	SVM	C4.5	SVM	C4.5
all-FeatureSet	MUC	52.95	58.60	59.92	59.48	**64.59**	59.23	63.33
	B^3	51.81	**84.20**	78.39	82.83	77.60	82.50	77.90
	CEAF	42.41	**78.02**	71.36	76.83	70.92	76.30	71.34
	BLANC	51.68	**66.93**	65.14	66.46	66.90	66.13	67.32
relational-FeatureSet	MUC	47.51	39.71	**50.69**	38.42	48.21	39.03	48.22
	B^3	34.32	76.25	**78.41**	75.55	32.18	75.87	32.31
	CEAF	29.91	63.57	**69.61**	62.76	27.83	62.22	27.86
	BLANC	47.40	61.43	**65.61**	60.71	23.06	60.10	24.55
notOral-FeatureSet	MUC	52.46	58.83	59.29	59.09	61.31	59.64	**63.37**
	B^3	55.16	**83.95**	79.38	82.63	77.95	82.35	78.62
	CEAF	44.83	**77.74**	72.39	76.57	71.28	76.03	72.04
	BLANC	54.60	66.44	65.47	66.19	**76.00**	65.69	67.00

Experiments and Results with ANCOR

Comments

- Naive Byes is not competitive
- small training sets are enough (over-fitting for larger training sets ?)
- SVM globally better
- use of all features useful
- oral-specific features are not very important (and not high in decision trees)
- important features (from decision trees) : distance between entities, equality of gender, NE nature of entities, embedding

Experiments and Results with ANCOR

Perspective

- learn separately for each sub-corpus
- feature selection by incremental adding of one feature after the other
- learn the typing of the links (or learn separately for distinct type)
- associate this system with an identification of entities : end-to-end system

Conclusion

- coreference resolution is difficult, there is room for improvements
- experimental study : a lot of parameters are to be taken into account (differences between various results are more important than the results themselves)
- machine learning does not prevent from thinking !