

# Evaluating Information Extraction

Andrea Esuli and Fabrizio Sebastiani

Istituto di Scienza e Tecnologie dell'Informazione  
Consiglio Nazionale delle Ricerche  
Via Giuseppe Moruzzi, 1 – 56124 Pisa, Italy  
E-mail: {*firstname.lastname*}@isti.cnr.it

Conference on  
Multilingual and Multimodal Information Access Evaluation  
(CLEF 2010)  
September 20-23, 2010 – Padova, IT

# (Annotation-based) Information Extraction: an example

The screenshot displays a software application for medical text annotation. The main window is titled "Document Editor" and contains a text area with a medical report. The text is annotated with colored boxes, and a checklist on the right side of the window shows the corresponding categories for each annotation. The checklist includes items such as "Esiti chirurgici", "BIRADS", "Enhancement descrizione", "Enhancement presenza/assenza", "Indicazioni Esame", "Informazioni Tecniche", "Linfonodi locoregionali", "Protesi descrizione", and "Terapie/follow-up". The text in the document editor is as follows:

Esame effettuato con magnete 1.5 T, mediante bobina dedicata, con sequenze T2 TIRM anche con soppressione del segnale dell'acqua e T1 flash 3D sui piani assiali e sagittale, prima e dopo la somministrazione di 0.1 mmol / L di mdc paramagnetico per via endovenosa. Sono state effettuate successivamente elaborazioni MIP e grafici dell'intensità del potenziamento. Indicazione all'esame: Controllo micondularita retroareolari a destra in paziente con pregressa QUART sinistra e mastoplastica additiva bilaterale. Sequenze T2 e T1 pesate precontrasto: Presenza di impianti protesici monocamera posizionati in sede retro ghiandolare bilateralmente. Entrambe le protesi risultano normoespanse, con regolari profili capsulari ed omogenea intensità di segnale del contenuto protesico. Non evidenza di segni di rottura intra od extracapsulare, nè falde fluide periprotetiche. A sinistra esiti di pregressa QUART. A destra il parenchima risulta localizzato prevalentemente in sede centrale e mostra regolare rappresentazione di entrambe le componenti ghiandolare e adiposa. A destra in corrispondenza della regione sovrareolare centrale si evidenzia piccola formazione nodulare solida (4 mm) a margini regolari iperintensa in T2, isolintensa nelle sequenze FS. Alcuni linfonodi di aspetto reattivo in entrambi i cavi ascellari (DM 7 mm). Sequenze T1 postcontrasto: A destra in sede sovrareolare centrale la formazione precedentemente descritta presenta precoce potenziamento a morfologia nodulare rotondeggiante a margini regolari con enhancement progressivo nel tempo ad andamento centrifugo. Le caratteristiche del potenziamento depongono per lesione verosimilmente benigna della mammella (BIRADS 3), meritevole di monitoraggio ecografico ravvicinato nel tempo (4 - 6 mesi). Non ulteriori aree di potenziamento sospette bilateralmente, nè significativi potenziamenti linfonodali. CODICE ACR : 00.7

The interface also features a file explorer on the left side, showing a list of files with their MIME types and document types. The file explorer shows the following files and their properties:

File Name	MimeType	docNewLineType	gate.SourceURL
464.xml_0011D	text/rt	CRLF	file:fil
463.xml_0011C			
460.xml_0011B			
452.xml_0011A			
451.xml_00119			
445.xml_00118			
444.xml_00117			
443.xml_00116			
442.xml_00115			
441.xml_00114			
440.xml_00113			
439.xml_00112			
432.xml_00111			
431.xml_00110			
430.xml_0010F			

The interface also features a toolbar at the top with various icons for file operations and editing. The bottom of the window shows the "Document Editor" and "Initialisation Parameters" tabs.

# Outline

- 1 Introduction
- 2 Defining Information Extraction
- 3 The Segmentation F-score
- 4 The Token & Separator Model
- 5 Experiments
- 6 Conclusion and further work

# Outline

- 1 Introduction
- 2 Defining Information Extraction
- 3 The Segmentation F-score
- 4 The Token & Separator Model
- 5 Experiments
- 6 Conclusion and further work

# Introduction

- Little past research and discussion on mathematical measures for evaluating **Information Extraction** (IE)
- Generalized feeling that no satisfactory measure has been found yet.
- The most frequently used evaluation model in IE is the **segmentation F-score**
- We claim that it suffers from several problems, and propose a new evaluation model that does not suffer from them.

# Introduction

- Little past research and discussion on mathematical measures for evaluating **Information Extraction** (IE)
- Generalized feeling that no satisfactory measure has been found yet.
- The most frequently used evaluation model in IE is the **segmentation F-score**
- We claim that it suffers from several problems, and propose a new evaluation model that does not suffer from them.

# Outline

- 1 Introduction
- 2 Defining Information Extraction**
- 3 The Segmentation F-score
- 4 The Token & Separator Model
- 5 Experiments
- 6 Conclusion and further work

# A formal definition of IE

- Let a text  $U = \{t_1 \prec s_1 \prec \dots \prec s_{n-1} \prec t_n\}$  consist of a sequence of
  - **tokens** (e.g., word occurrences)  $t_1, \dots, t_n$  and
  - **separators** (e.g., sequences of blanks and punctuation symbols)  $s_1 \dots s_{n-1}$

The term **textual unit** (or simply **t-unit**) denotes either a token or a separator.

- Let  $C = \{c_1, \dots, c_m\}$  be a predefined set of **tags**, or **tagset**.
- Let  $A = \{\sigma_{11}, \dots, \sigma_{1k_1}, \dots, \sigma_{m1}, \dots, \sigma_{mk_m}\}$  be an **annotation** for  $U$ , where a **segment**  $\sigma_{ij}$  for  $U$  is a pair  $(st_{ij}, et_{ij})$  composed of a start token  $st_{ij} \in U$  and an end token  $et_{ij} \in U$ .



# A formal definition of IE (cont'd)

- We define **Information Extraction (IE)** as the task of estimating an unknown target function  $\Phi : \mathcal{U} \times \mathcal{C} \rightarrow \mathcal{A}$ , that defines how a text  $U \in \mathcal{U}$  ought to be annotated (according to a tagset  $\mathcal{C}$ ) by an annotation  $A \in \mathcal{A}$ . The result  $\hat{\Phi} : \mathcal{U} \times \mathcal{C} \rightarrow \mathcal{A}$  of this estimation is called a **tagger**.
- Given
  - a **true annotation**  $A = \Phi(U, C) = \{\sigma_{11}, \dots, \sigma_{1k_1}, \dots, \sigma_{m1}, \dots, \sigma_{mk_m}\}$
  - a **predicted annotation**  $\hat{A} = \hat{\Phi}(U, C) = \{\hat{\sigma}_{11}, \dots, \hat{\sigma}_{1\hat{k}_1}, \dots, \hat{\sigma}_{m1}, \dots, \hat{\sigma}_{m\hat{k}_m}\}$
 our aim is that of defining precise criteria for measuring how accurate this estimation is.

# Single-tag IE or Multi-tag IE?

- Our definition allows a given t-unit to be tagged by more than one tag (**multi-tag IE**).
  - Example: in the expression “the Ronald Reagan Presidential Library” we might decree the t-units in “Ronald Reagan” to be instances of both the PER (“person”) tag and the ORG (“organization”)
- **Single-tag IE** is a special case of multi-tag IE, and a measure for multi-tag IE by definition accounts for single-tag IE too.
- Multi-tag IE thus consists of  $m$  independent subproblems of estimating  $\hat{\Phi}_i : \mathcal{U} \rightarrow \mathcal{A}_i$ , for any  $i \in \{1, \dots, m\}$ . We will thus simply deal with  **$c_i$ -annotations**, i.e., sets of  **$c_i$ -segments** of the form  $A_i = \{\sigma_{i1}, \dots, \sigma_{ik_i}\}$ , for any  $i \in \{1, \dots, m\}$ .

# Outline

- 1 Introduction
- 2 Defining Information Extraction
- 3 The Segmentation F-score**
- 4 The Token & Separator Model
- 5 Experiments
- 6 Conclusion and further work

# The segmentation F-score

## Example

<i>true</i>	The <u>quick brown</u> fox jumps over the <u>lazy dog</u>	FN	FN
<i>predicted</i>	The <u>quick brown fox</u> jumps over the <u>lazy dog</u>	FP	FP FP

- The segmentation F-score model assumes
  - 1 IE to be a **single-tag** task
  - 2  $F_1 = \frac{2TP}{FP + FN + 2TP}$  as **the evaluation measure**
  - 3 The set of segments (true or predicted) as **the event space**
- These choices give rise to problems

# Problems with the segmentation F-score: 1. True negatives

- Assumption 3 makes the notion of a true negative (“any segment of any length that is neither a true nor a predicted segment”) too clumsy to be of any real use.
- There are  $O(n^2)$  such TNs ...
- While this is not a problem for  $F_1$ , this would not allow switching to other plausible measures of agreement (e.g., Cohen’s kappa, ROC, accuracy).

## Problems with the segmentation F-score: 2. Overlap

- In the segmentation F-score there are several alternative models of what counts as a TP:
  - **Exact match model** (most frequently used one): only exact matches count as TPs;
    - too harsh (e.g., for tag ORG,  $\sigma$  = "Ronald Reagan Presidential Library",  $\hat{\sigma}$  = "Reagan Presidential Library" count as a double mistake, since  $\sigma$  is a FN and  $\hat{\sigma}$  is a FP);
  - **Overlap model**: if  $\sigma$  and  $\hat{\sigma}$  overlap even marginally, this is a TP:
    - too lenient
    - encourages "cheating" (e.g., when  $\hat{\sigma}$  covers the entire document ...)
  - **Constrained overlap model**: max  $k_1$  spurious tokens and max  $k_2$  missing tokens are accepted:
    - too arbitrary;
    - does not reward exact matches (e.g.,  $\hat{\sigma}'$  = "the Ronald Reagan Presidential" is given the same credit as  $\hat{\sigma}''$  = "Ronald Reagan Presidential Library")

## Problems with the segmentation F-score: 3. Tag switches

- Not clear how to deal with **tag switches**, i.e., with cases in which the boundaries of a segment have been recognized (more or less exactly, according to one of the three models above) but the right tag has not.
- E.g., tagging “San Diego” as PER instead of LOC

# Outline

- 1 Introduction
- 2 Defining Information Extraction
- 3 The Segmentation F-score
- 4 The Token & Separator Model**
- 5 Experiments
- 6 Conclusion and further work



# The Token & Separator Model

- The solution we propose is based on using the set of all t-units as the event space; we dub it **the Token & Separator Model** (or **TS model**).

## Example

<i>true</i>	The	○	quick	○	brown	○	fox	○	jumps	○	over	○	the	○	lazy	○	dog
<i>predicted</i>	The	○	quick	○	brown	○	fox	○	jumps	○	over	○	the	○	lazy	○	dog
	TN	TN	TP	TP	TP	FP	FP	TN	TN	TN	TN	TN	TN	TN	TP	FN	TP

This example returns the following scores:

Segmentation F-score with exact match  $F_1 = 0$

Segmentation F-score with overlap match  $F_1 = 1$

TS model (with  $F_1$ )

$$F_1 = \frac{2 * 5}{2 * 5 + 2 + 1} = .77$$

# The Token & Separator Model (cont'd)

- The TS model addresses the three shortcomings of the segmentation F-score:
  - ① The TS model contemplates “reasonable” true negatives
  - ② The TS model naturally accounts for degree of overlap, with no need for numerical parameters
  - ③ The TS model naturally deals with tag switches, since each tag is addressed separately

# The Token & Separator Model (cont'd)

- Separators are included in the event space so as to correctly evaluate segment boundary recognition: e.g., assume we need to extract PER from “Barack Obama, Hillary Clinton and Joe Biden” ...

## Example

<i>true</i>	Barack	○	Obama	○	Hillary	○	Clinton	○	and	○	Joe	○	Biden
<i>predicted</i>	Barack	○	Obama	○	Hillary	○	Clinton	○	and	○	Joe	○	Biden
	TP		TP		TP		FP		TP		TP		TP

## The Token & Separator Model (cont'd)

- $F_1$  or other (e.g., Cohen's kappa) may be used as the measure; macro- or micro- may be used as the averaging method.
- Sticking to  $F_1$  as the measure has several advantages:
  - robust to high imbalance;
  - does not encourage a tagger to either undertag or overtag;
  - may be modified (as  $F_\beta$ ) to accommodate higher penalty for overtagging or undertagging;
  - learning algorithms for IE that are capable of internally optimizing  $F_1$  are available.
- Adopting macro- as the averaging method has also advantages:
  - Does not reward systems only good at tagging frequent tags

# Outline

- 1 Introduction
- 2 Defining Information Extraction
- 3 The Segmentation F-score
- 4 The Token & Separator Model
- 5 Experiments**
- 6 Conclusion and further work

# Experiments

- We have re-evaluated according to the  $TS-F_1^M$  model the submissions to the CoNLL'02 and CoNLL'03 NER Shared Tasks. Original evaluation was performed with segmentation F-score and exact match.
  - CoNLL'02: 12 participants, Spanish and Dutch NER (we could not reevaluate Dutch since original files no longer available).
  - CoNLL'03: 16 participants, English and German NER.

SPANISH	Seg- $F_1$	1	2	3	4	5	6	7	8	9	10	11	12				
		.814	.791	.771	.766	.758	.758	.739	.739	.737	.715	.637	.610				
	$TS-F_1^M$	1	2	4	5	6	7	10	9	8	3	12	11				
		.821	.799	.769	.746	.746	.740	.734	.729	.724	.710	.677	.636				
				-7	+1	+1	+1	+1	-1	+1	+3	-1	+1				
ENGLISH	Seg- $F_1$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
		.888	.883	.861	.855	.850	.849	.847	.843	.840	.839	.825	.817	.798	.782	.770	.602
	$TS-F_1^M$	1	2	3	4	11	8	6	5	10	7	9	14	15	13	12	16
		.875	.874	.857	.853	.848	.845	.842	.840	.835	.833	.819	.817	.813	.809	.808	.671
						-3	-1	-3	+2	-2	+1	+6	-3	-1	+2	+2	
GERMAN	Seg- $F_1$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
		.724	.719	.713	.700	.692	.689	.684	.681	.678	.665	.663	.657	.630	.573	.544	.477
	$TS-F_1^M$	1	9	3	2	4	7	6	5	8	11	10	13	12	14	15	16
		.719	.708	.706	.702	.695	.691	.690	.679	.674	.650	.645	.642	.641	.616	.569	.471
			-2		-1	-3	-1	+1	-1	+7	-1	+1	-1	+1			

## Experiments: Anecdotal evaluation

- The participant that placed 3rd in CoNLL'02 Spanish is ranked 10th (3rd from last!) by the TS model.
- The participant that placed 11th in CoNLL'03 English is ranked 5th by the TS model. With respect to the participant that placed 5th
  - - It generated 2% fewer exact matches
  - + It generated 158% more “close matches” – i.e., accurate modulo a single token
  - + It totally missed 7% fewer segments
- The participant that placed 9th in CoNLL'03 German is ranked 2nd by the TS model.
- **Taking a stand between the two models is important!**

## Experiments: Rank correlation

- We have computed the Spearman's rank correlation

$$R(\eta', \eta'') = 1 - \frac{6 \sum_{k=1}^p (\eta'(\hat{\Phi}_k) - \eta''(\hat{\Phi}_k))^2}{p(p^2 - 1)}$$

(averaged across the English, German, and Spanish tasks) between the results produced by the different evaluation models.

$R(\eta', \eta'')$	Seg- $F_1$	TS- $F_1^M$	T- $F_1^M$
Seg- $F_1$	1.0	.832	.832
TS- $F_1^M$	.832	1.0	.990
T- $F_1^M$	.832	.990	1.0



# Outline

- 1 Introduction
- 2 Defining Information Extraction
- 3 The Segmentation F-score
- 4 The Token & Separator Model
- 5 Experiments
- 6 Conclusion and further work**

## Conclusion and further work

- Overcome shortcomings of segmentation F-score by
  - clearly separating event space and evaluation measure
  - using the set of tokens and separators as the former.
- Scorer for IOB2 format available at <http://patty.isti.cnr.it/esuli/IEevaluation/> (computes both segmentation F-score and  $TS-F_1^M$  model).
- Problem: The TS model does not work for **multi-instance IE** (i.e., when the same token/separator may belong to more than one segment for the same tag – as e.g., in opinion extraction under the WWC tagset).

# The TS model: Potential criticisms

- **Q:** My IE application is actually single-tag, and the TS model was developed for multi-tag IE ...
- **A:** Single-tag is a special case of multi-tag. If the true annotation is single-tag, our evaluation model indeed penalizes a tagger for not generating a single-tag prediction.
- The same goes for **single-segment IE** ...

## The TS model: Potential criticisms (cont'd)

- **Q:** The TS model wrongly treats all tokens (e.g., articles and nouns) as having equal importance ...
- **A:** If desired, different weights may be assigned to individual tokens/separators in the true annotation, since most contingency-table-based measures may be extended to deal with “weighted events”.

## The TS model: Potential criticisms (cont'd)

- **Q:** The TS model places too much importance on separators ...
- **A:** Again, different weights may be assigned to tokens and separators, if desired, in the true annotation. Anyway,  $R(\eta', \eta'') = .990$  shows that rankings are not modified substantially even by completely removing separators from consideration.

## The TS model: Potential criticisms (cont'd)

- **Q:** Is the TS model too harsh on tag switches? E.g., system that correctly identifies the boundaries of segment “San Diego” but incorrectly tags it as PER instead of LOC, is assigned three FNs for LOC and three FPs for PER.
- **A:**
  - Not too severe a penalty in the general case in which the two tags are not known to be close in meaning.
  - When tags are known to be close in meaning (e.g., PER, LOC, ORG, MISC), a common supertag (“NE”) may be created and evaluation may also be carried out in terms of it.