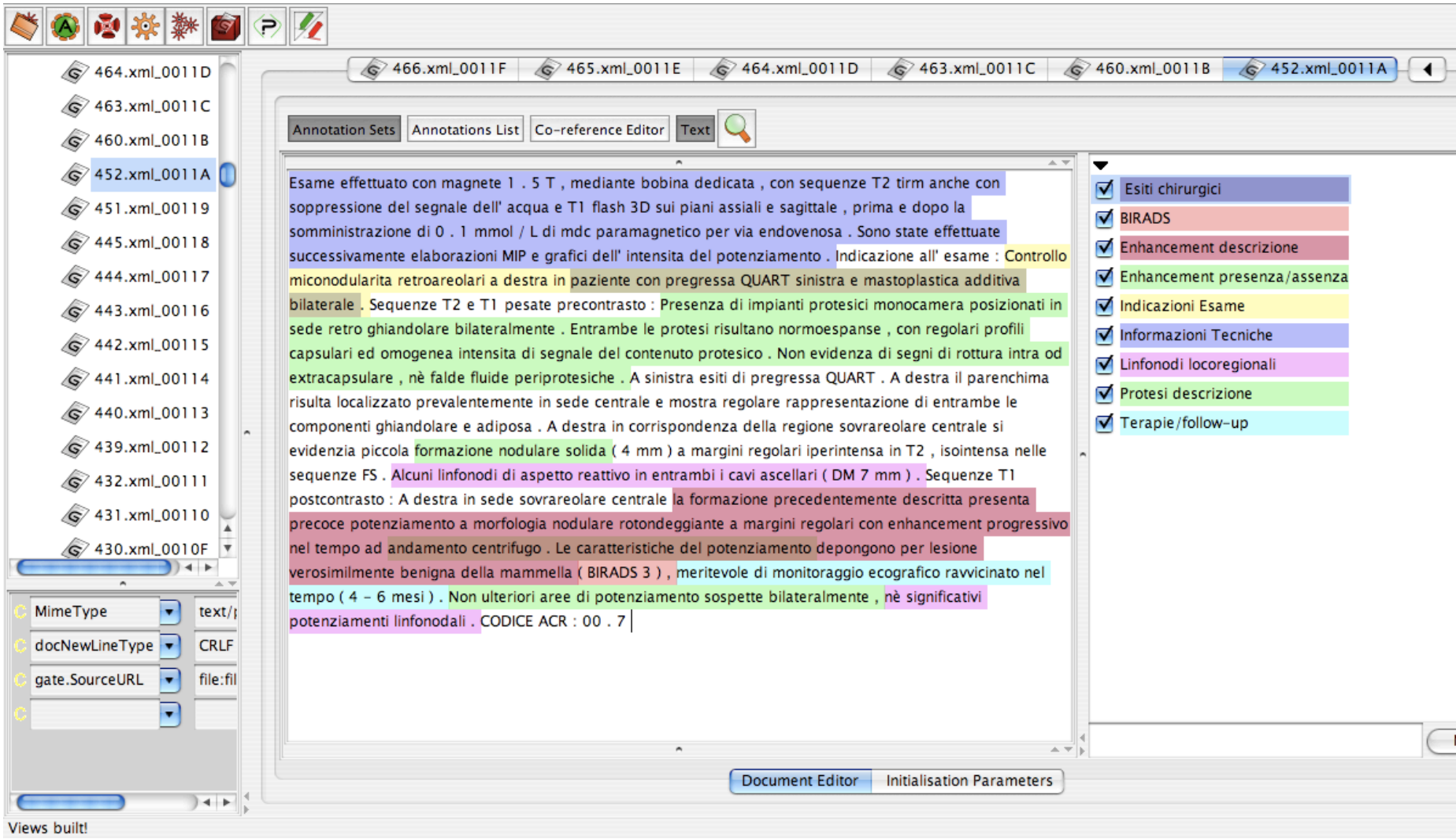
Evaluating Information Extraction

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



Outline

Outline

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

A formal definition of IE

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

  The term textual unit (or simply t-unit) denotes either a token or a separator.
- Let $C = \{c_1, \ldots, c_m\}$ be a predefined set of tags, or tagset.
- Let $A = \{\sigma_{11}, \ldots, \sigma_{1k_1}, \ldots, \sigma_{m1}, \ldots, \sigma_{mk_m}\}$ be an annotation for U, where a segment σ_{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 $\Phi : \mathcal{U} \times C \rightarrow \mathcal{A}$, that defines how a text $U \in \mathcal{U}$ ought to be annotated (according to a tagset C) by an annotation $A \in \mathcal{A}$. The result $\hat{\Phi} : \mathcal{U} \times C \rightarrow \mathcal{A}$ of this estimation is called a tagger.
- Given
  a true annotation $A = \Phi(U, C) = \{\sigma_{11}, \ldots, \sigma_{1k_1}, \ldots, \sigma_{m1}, \ldots, \sigma_{mk_m}\}$
  a predicted annotation $\hat{A} = \hat{\Phi}(U, C) = \{\hat{\sigma}_{11}, \ldots, \hat{\sigma}_{1\hat{k}_1}, \ldots, \hat{\sigma}_{m1}, \ldots, \hat{\sigma}_{m\hat{k}_m}\}$
  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 $i \in \{1, \ldots, m\}$. We will thus simply deal with c_i-annotations, i.e., sets of c_i-segments of the form $A_i = \{\sigma_{i1}, \ldots, \sigma_{ik_i}\}$, for any $i \in \{1, \ldots, m\}$.

Outline

The segmentation F-score

Example

		FN					FN	
true	The	quick	brown	fox	jumps	over	the	lazy dog
predicted	The	quick	brown	fox	jumps	over	the	lazy dog
		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, σ="Ronald Reagan Presidential Library", $\hat{\sigma}$="Reagan Presidential Library" count as a double mistake, since σ is a FN and $\hat{\sigma}$ is a FP);
  - Overlap model: if σ 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

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

true	The ◯ quick ◯ brown ◯ fox ◯ jumps ◯ over ◯ the ◯ lazy ◯ dog
predicted	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

true	Barack	◯	Obama	◯	Hillary	◯	Clinton	◯	and	◯	Joe	◯	Biden
predicted	Barack	◯	Obama	◯	Hillary	◯	Clinton	◯	and	◯	Joe	◯	Biden
	TP		TP		TP		FP		TP		TP		TP		TN		TN		TN		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_β) 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

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

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.