Evaluating Information Extraction

Andrea Esuli
Fabrizio Sebastiani
Consiglio Nazionale delle Ricerche,
Italy*

The issue of how to experimentally evaluate information extraction (IE) systems has received hardly any satisfactory solution in the literature. In this paper we propose a novel evaluation model for IE and argue that, among others, it allows (i) a correct appreciation of the degree of overlap between predicted and true segments, and (ii) a fair evaluation of the ability of a system to correctly identify segment boundaries. We describe the properties of this models, also by presenting the result of a re-evaluation of the results of the CoNLL'03 and CoNLL'02 Shared Tasks on Named Entity Extraction.

1. Introduction

The issue of how to measure the effectiveness of information extraction (IE) systems has received little attention, and hardly any definitive answer, in the literature. A recent review paper on the evaluation of IE systems (Lavelli et al. 2008), while discussing in detail other undoubtedly important evaluation issues (such as datasets, training set / test set splits, and evaluation campaigns), devotes surprisingly little space to discussing the mathematical *measures* used in evaluating IE systems; and the same happens for a recent survey on information extraction methods and systems (Sarawagi 2008). That the issue is far from solved is witnessed by a long discussion[1], appeared on a popular NLP-related blog, in which prominent members of the NLP community voice their discontent with the evaluation measures currently used in the IE literature, and come to the conclusion that no satisfactory measure has been found yet.

The lack of agreement on an evaluation measure for IE has several negative consequences. The first is that we do not have an agreed way to compare different IE techniques on shared benchmarks, which in itself is a hindrance to the progress of the discipline. The second is that, since IE is usually tackled via machine learning techniques, we do not have an agreed measure that learning algorithms based on explicit loss minimization can optimize. The third is that, whenever we optimize the parameters of our favourite IE technique via cross-validation, we generate parameter choices that are optimal for an evaluation measure of dubious standing.

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

Submission received: February 26, 2010

1 Christopher Manning, Hal Daume III, and others, *Doing Named Entity Recognition? Don't optimize for F_1*, `http://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html`, accessed on February 26, 2010. The discussion is actually framed in terms of evaluating *named entity recognition* (NER), but all of it straightforwardly applies to IE tasks other than NER.

A favourite measure for evaluating IE systems is F_1 (Lewis 1995; van Rijsbergen 1974), defined as the harmonic mean of the well-known notions of *precision* (π) and *recall* (ρ):

$$F_1 = \frac{2\pi\rho}{\pi+\rho} = \frac{2\frac{TP}{TP+FP}\frac{TP}{TP+FN}}{\frac{TP}{TP+FP}+\frac{TP}{TP+FN}} = \frac{2TP}{FP+FN+2TP} \quad (1)$$

In the IE incarnation of F_1, the symbols TP, FP, and FN stand for the numbers of true positives, false positives, and false negatives, resulting from a standard binary contingency table computed on the true and predicted "segments", where a segment is taken to be correctly recognized only when its boundaries have been exactly identified. As a result, this evaluation model is sometimes called *segmentation F-score* (Suzuki, McDermott, and Isozaki 2006). In this paper we argue that the segmentation F-score model has several shortcomings, and propose a new evaluation model that does not suffer from them.

The rest of the paper is organized as follows. Section 2 gives preliminary definitions. Section 3 discusses the shortcomings of the segmentation F-score model in detail, while Section 4 goes on to present our alternative model. In Section 5 we re-evaluate a number of past experiments from the literature in terms of our proposed model, and show that the two models rank competing systems in a substantively different way. Section 6 concludes by sketching avenues for future work.

2. A formal definition of information extraction

Let a text $U = \{t_1 \prec s_1 \prec \ldots \prec s_{n-1} \prec t_n\}$ consist of a sequence of *tokens* (typically: word occurrences) $t_1, \ldots, t_n$ and *separators* (typically: sequences of blanks and punctuation symbols) $s_1, \ldots, s_{n-1}$, where "$\prec$" means "precedes in the text". We use the term *textual unit* (or simply *t-unit*), with variables $u_1, u_2, \ldots$, to denote either a token or a separator. Let $C = \{c_1, \ldots, c_m\}$ be a predefined set of *tags* (aka *labels*, or *classes*), 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* $et_{ij} \in U$ such that $st_{ij} \preceq et_{ij}$ ("$\preceq$" obviously means "either precedes in the text or coincides with"). Here, the intended semantics is that, given segment $\sigma_{ij} = (st_{ij}, et_{ij}) \in A$, all t-units between st_{ij} and et_{ij}, extremes included, are tagged with tag c_i[2].

Given a universe of texts $\mathcal{U}$ and a universe of segments $\mathcal{A}$, we define *information extraction* (IE) as the task of estimating an unknown target function $\Phi : \mathcal{U} \times C \to \mathcal{A}$, 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 $\hat{\Phi} : \mathcal{U} \times C \to \mathcal{A}$ of this estimation is called a *tagger*. Our aim in this paper is exactly that of defining precise criteria for measuring how accurate this estimation is[3].

Given a *true annotation* $A = \Phi(U, C) = \{\sigma_{11}, \ldots, \sigma_{1k_1}, \ldots, \sigma_{m1}, \ldots, \sigma_{mk_m}\}$ and a *predicted annotation* $\hat{A} = \hat{\Phi}(U, C) = \{\hat{\sigma}_{11}, \ldots, \hat{\sigma}_{1\hat{k}_1}, \ldots, \hat{\sigma}_{m1}, \ldots, \hat{\sigma}_{m\hat{k}_m}\}$, for any $i \in \{1, \ldots, m\}$ we naturally make the general assumption that $\hat{k}_i$ may differ from k_i; that

2 The reason why also the separators are the objects of annotation will become apparent in Section 4.
3 Consistently with most mathematical literature we use the caret symbol (ˆ) to indicate estimation.

is, a tagger may in general produce, for a given tag c_i, more segments that it should, or less segments than it should.

The notion of IE we have defined allows in principle a given t-unit to be tagged by more than one tag, and might thus be dubbed *multi-tag IE*. A specific real application, also depending on the tagset considered, might have a multi-tag nature or not. For instance, 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") tag; or we might decree them to be only instances of the ORG tag. The aim of the present paper is to propose an evaluation model for IE that is intuitive and plausible irrespectively of whether the applications we are dealing with have a single-tag or a multi-tag nature.

While different IE applications might want to take different stands on the single-tag vs. multi-tag issue, it is important to note that our definition above is general, since single-tag IE is just a special case of multi-tag IE. If the true set of segments is single-tag, it will be the task of the tagger to generate a single-tag prediction, and it will be the task of the evaluation model to penalize a tagger for not doing so. The multi-tag nature of our definition essentially means that, given tagset $C = \{c_1, \ldots, c_m\}$, we can split our original problem into m independent subproblems of estimating a target function $\Phi_i : \mathcal{U} \rightarrow \mathcal{A}_i$ by means of a tagger $\hat{\Phi}_i : \mathcal{U} \rightarrow \mathcal{A}_i$, for any $i \in \{1, \ldots, m\}$. Likewise, the annotations we will be concerned with from now on will actually be c_i*-annotations*, i.e., sets of c_i*-segments* of the form $A_i = \{\sigma_{i1}, \ldots, \sigma_{ik_i}\}$.

3. Problems with the current evaluation model

Our proposal for evaluating IE is based on carefully distinguishing the *mathematical measure* to be adopted for evaluation from the *event space* (i.e., universe of objects) to which this measure is applied. From this point of view, we have seen in Section 1 that the standard "segmentation F-score" model of evaluating IE systems assumes F_1 as the evaluation measure and the set of segments (true or predicted) as the event space. However, this particular choice of event space is problematic.

One problem is that the choice of segments as the event space makes the notion of a "true negative" too clumsy to be of any real use: a true negative should be a sequence (of any length) of tokens and separators that is neither a true nor a predicted segment, and the number of such sequences in a text of even modest length is combinatorially large, and simply too large to be of any use. While this does not prevent F_1 from being used as a measure, since F_1 is not a function of the number of true negatives (see Equation 1), this would not allow the use of other plausible measures of agreement between true and predicted annotation (such as e.g., Cohen's kappa (Cohen 1960), ROC analysis (Fawcett 2006), or simple accuracy) that are indeed a function of the number of true negatives.

A second problem is that it is not clear how partial overlap should be treated. While a true segment that perfectly coincides with a predicted segment is no doubt a true positive, when should a true segment that *partially* coincides with a predicted segment be treated as a true positive?

According to the *exact match model* (currently the most frequently used model; see e.g., (Freitag and Kushmerick 2000; Freitag 2000; Krishnan and Manning 2006; Tjong Kim Sang 2002; Tjong Kim Sang and De Meulder 2003; Suzuki, McDermott, and Isozaki 2006)) this should never be the case. This seems too harsh a criterion: for instance, given true segment σ="Ronald Reagan Presidential Library" for tag ORG, a tagger that tags as ORG the segment $\hat{\sigma}$="Reagan Presidential Library" would receive no credit at all for

this (σ would generate a false negative and $\hat{\sigma}$ would generate a false positive). Even worse, this tagger would receive even less credit than a tagger that predicts no segment overlapping with σ (this would generate a false negative but no false positive).

Conversely, the (less frequently used) *overlap* model (Freitag 1997) returns a true positive whenever the tagger predicts a segment $\hat{\sigma}$ that overlaps even marginally with the true segment σ. This seems too lenient a criterion; in the extreme, a tagger that generates a single segment that covers the entire text U would obtain a perfect score, since every true segment overlaps with the single predicted segment.

A more sophisticated variant is what we might call the *constrained overlap* model (De Sitter and Daelemans 2003), in which only overlaps with at most k_1 spurious tokens and at most k_2 missing tokens are accepted as valid. This model, while less lenient, is problematic because of its dependence on parameters (k_1 and k_2), since any choice of actual values for them may be considered arbitrary. Additionally, this model does not adequately reward taggers that identify the boundaries of a segment exactly; for instance, given true segment σ="Ronald Reagan Presidential Library" for tag ORG, and given parameter choices $k_1 = 1$ and $k_2 = 1$, a tagger that tags as ORG the segment $\hat{\sigma}'$="the Ronald Reagan Presidential" is given the same credit as one that instead returns $\hat{\sigma}''$="Ronald Reagan Presidential Library". Similar drawbacks are presented by the *contain* model (Freitag 1997), which is actually a special case of the constrained overlap model in which $k_2 = 0$, and by variants of these models that have been proposed for the specific needs of biomedical NER (Tsai et al. 2006).

A third problem is that, when F_1 is used as the evaluation measure and the set of segments is used as the event space, it is 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 four models above) but the right tag has not (e.g., when a named entity has been correctly recognized as such but it has been incorrectly deemed as one of type PER instead of type ORG). The problems of partial overlap and tag switch may of course nastily interact, just adding to the headache.

4. The token & separator F_1^M model

Essentially, the analysis of the existing IE evaluation model(s) that we have carried out in the previous section indicates that a new, improved model should (i) allow in a natural way for the notion of a "true negative", (ii) be sensitive to the *degree* of overlap between true and predicted segments, and (iii) naturally model "tag switches" and the problems arising from the presence of multiple tags in a given tagset.

4.1 The event space

The solution we propose is based on using *the set of all tokens and separators (i.e., the set of all t-units) as the event space*; we dub it *the token & separator model* (or *TS model*). In this solution, desideratum (i) is achieved by having true negatives consist of simple t-units, and not combinations of them; this has the advantage of being a more natural choice, and of bounding the number of true negatives by the number of t-units in the text. Desideratum (ii) is instead achieved by making the analysis more granular, and making a (true or predicted) segment contribute not one but *several* units to the contingency table, proportionally to its length. As for desideratum (iii), we will discuss how it is achieved later on in this section.

Let us assume for a moment that we stick to F_1 as the evaluation measure and that our tagset contains a single tag c_i, and let us look at the example annotated sentence of

Table 1
Example sentence annotated according to a single tag c_i; A_i is the true annotation while $\hat{A}_i$ is the predicted annotation. For higher readability all tokens and separators (blanks, in this case) are numbered from 1 to 17.

A_i			c_i	c_i	c_i										c_i	c_i	c_i
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	The		quick		brown		fox		jumps		over		the		lazy		dog
$\hat{A}_i$			c_i	c_i	c_i	c_i	c_i								c_i		c_i

Table 1. For this example we have $F_1 = \frac{2TP}{2TP+FP+FN} = \frac{2*5}{2*5+2+1} = .769$, as deriving from the presence of 5 true positives (tokens "quick" and "brown" and the separator between them, plus tokens "lazy" and "dog"), 2 false positives (token "fox" and the separator before it) and 1 false negative (the separator between "lazy" and "dog"). The same example would have resulted in $F_1 = 0$ under the exact match model (since no segment is perfectly recognized) and $F_1 = 1$ under the overlap model (since all segments are at least partially recognized). The results according to the other two models would obviously depend on the parameter choices for k_1 and k_2.

The TS model finally makes it clear why, in the definitions of Section 2, we consider separators to be the object of tagging too: the reason is that the IE system should correctly identify segment boundaries. For instance, given the expression "Barack Obama, Hillary Clinton and Joe Biden" the perfect IE system will attribute the PER tag, among others, to the tokens "Barack", "Obama", "Hillary", "Clinton", and to the separators (in this case: blank spaces) between "Barack" and "Obama" and between "Hillary" and "Clinton", but *not* to the separator ", " between "Obama" and "Hillary". If the IE system does so, this means that it has correctly identified the boundaries of the segments "Barack Obama" and "Hillary Clinton". In the example annotated sentence of Table 1, the imperfect extraction of the second segment "lazy dog" via the two subsegments "lazy" and "dog" results in two true positives and one false negative; in this case, the system is partially penalized for having failed to recognize that "lazy" and "dog" are not two separate segments, and that together they form a unique segment.

By moving from a model of events as segments to a more granular model of events as tokens and separators, this model thus takes in the right account the degree of overlap of true and predicted segments, and does so without resorting to numerical parameters that would require arbitrary decisions for their setting. Furthermore, by taking also separators into account, it correctly distinguishes the case of consecutive but separate segments, from the case of a single long segment consisting of their concatenation.

Note also that, in the TS model, the cardinality of the set of events (i.e., the total of the four figures in the contingency table) is fixed, since it coincides with the length $L(U) = 2n - 1$ of text U (where "length" here also takes separators into account). This is in sharp contrast with the usual model in which segments are events, since in the latter model the cardinality of the set of events depends on the prediction (i.e., a predicted annotation $\hat{A}$ that contains many segments will generate sets of events with high cardinality, and vice versa). The result of this move is that different predicted annotations are now compared with reference to the *same* contingency table, and not to different contingency tables, which is fairly foreign to the tradition of contingency-table-based evaluations.

Let us now discuss desideratum (iii) above by examining the case of a tagset $C = \{c_1, \ldots, c_m\}$ consisting of more than one tag (for the moment being let us still stick to F_1 as the evaluation measure). The TS model is naturally extended to this case by viewing the task of annotating U according to the m tags, as consisting of m essentially independent tasks. As a result, evaluation can be carried out by computing m separate contingency tables for the m individual tags $c_i \in C$, and averaging the results across the tags.

Borrowing from the tradition of information retrieval evaluation, we can either adopt *microaveraged* F_1 (denoted by F_1^μ) or *macroaveraged* F_1 (F_1^M). F_1^μ is obtained by (i) computing the category-specific values TP_i, FP_i and FN_i, (ii) obtaining TP as the sum of the TP_i's (same for FP and FN), and then (iii) applying Equation (1). F_1^M is instead obtained by first computing the category-specific F_1 values and then averaging them across the c_i's.

It is well-known (see e.g., (Sebastiani 2002)) that, for the same annotation, F_1^μ and F_1^M may yield very different results. In fact, F_1^μ tends to be heavily influenced by the results obtained for the more frequent tags, since for these tags TN, FN and FP (the only arguments of the F_1 function) tend to be higher than for the infrequent tags. F_1^M has instead a more "democratic" character, since it gives the same importance to every tag in the tagset. As a result, it tends to return lower values than the (somehow overoptimistic) ones returned by F_1^μ, and to reward the systems that behave well also on the more infrequent tags. Because of this important property, we propose the adoption of *macroaveraging* as the default way of averaging results across tags.

A potential criticism of the fact that tagging under tagset $C = \{c_1, \ldots, c_m\}$ is evaluated as consisting of m independent tasks, is that certain tag switches may result in too severe a penalty. For instance, a system that correctly identifies the boundaries of segment "San Diego" but incorrectly tags it as PER instead of LOC is assigned three false negatives (for failing to recognize the LOC character of the segment) *and* three false positives (for incorrectly deeming the segment an instance of PER). We feel that this is actually not too severe a penalty in the general case in which the two involved tags are not known to be close in meaning. For instance, in an opinion extraction task (see e.g., (Wiebe, Wilson, and Cardie 2005)), the AGENT tag (that denotes either the source or the target agent of a "private state") and the DIRECT-SUBJECTIVE tag (that denotes either the explicit mention of a private state or a speech event expressing a private state) denote two concepts very distant in meaning, so distant that it seems reasonable to evaluate a tag switch between them as involving *both* false positives and false negatives. Conversely, in a task such as NER in which the different tags (PER, LOC, ORG, MISC) are close in meaning, the tags may be viewed as subtags of a common supertag ("ENTITY"). If desired, a more lenient evaluation may be performed by also evaluating ENTITY as a tag in its own. At this less granular level, correctly identifying the boundaries of a LOC segment but mistagging it as PER, would only give rise to true positives; this would provide, when desired, a coarser level of analysis that is more lenient towards tag switches between semantically related tags.

4.2 The evaluation measure

Concerning the evaluation measure to adopt, it is interesting to see that, in combination with the TS model, the problems that had plagued F_1 (and that had prompted "Don't optimize for F_1!" recommendations – see Footnote 1) disappear, which makes F_1 a plausible evaluation measure for IE. Concerning this, an interesting property of F_1 is that it does not depend on true negatives, which are going to be in very high numbers in

many IE applications such as NER; in other words, F_1 is inherently robust to the typical high imbalance between the positive and the negative examples of a tag. A second interesting property of F_1 is that it does not encourage a tagger to either undertag or overtag, since the trivial rejector (i.e., the tagger that does not tag any t-unit) has an F_1 score of 0, and the trivial acceptor (i.e., the tagger that tags all t-units) has an F_1 score equal to the fraction of true tagged t-units, which is usually very low. A third useful property of F_1 is that its more general form ($F_\beta = \frac{(\beta^2+1)\pi\rho}{\beta^2\pi+\rho} = \frac{(\beta^2+1)TP}{TP+FP+\beta^2(TP+FN)}$ – see e.g., (Lewis 1995)) also allows, if needed, a higher penalty to be placed on overtagging than undertagging (this is accomplished by picking a value of β in $[0, 1)$, with lower values placing heavier penalties) or viceversa (β in $(1, +\infty)$, with higher values placing heavier penalties). Last, it should be mentioned that learning algorithms for IE that are capable of internally optimizing for F_1 are available (in both the support vector machines camp – see (Joachims 2005) – and the conditional random fields camp – see (Suzuki, McDermott, and Isozaki 2006)), thus making it possible to generate taggers that are accurate at maximizing the two factors that our TS model rewards, i.e., (i) the degree of overlap between true and predicted segments, and (ii) the ability to correctly identify segment boundaries.

Given the fact that we advocate using (a) the set of tokens and separators as the set of events, (b) F_1 as the evaluation function, and (c) macroaveraging as the method for averaging results across tags, we will henceforth refer to our proposed model as *the token & separator F_1^M model* (or *TS-F_1^M model*).

5. Experiments

In order to provide an indication of the impact that our proposed model may have on a concrete evaluation, we have re-evaluated according to the TS-F_1^M model the submissions to the CoNLL'03 (Tjong Kim Sang and De Meulder 2003)[4] and CoNLL'02 (Tjong Kim Sang 2002)[5] Named Entity Extraction Shared Tasks. The CoNLL'03 NER Shared Task attracted 16 participants, and consisted of two subtasks, one on English and the other on German NER. The CoNLL'02 NER Shared Task attracted instead 12 participants, who competed on both Spanish and Dutch NER. We here deal only with the 2003 English and German data and with the 2002 Spanish data; we could not re-evaluate the 2002 Dutch data since the original files are no longer available due to copyright problems[6].

The 1st row of Table 2 presents the way the 16 participants on 2003 English data are ranked according to the segmentation F-score ("segment-based, exact-match F_1^μ model", in our terminology) officially adopted in the shared task, while the 2nd row reports the same for the TS-F_1^M model. Although the two rankings are not too dissimilar (e.g., the first 4 positions are the same), there are a few relevant differences. The participant that originally placed 11th in CoNLL'03 is ranked in 5th position by our evaluation model, jumping no less than 6 positions up in the ranking. This indicates that the algorithm of the 11th participant was perhaps suboptimal at producing exact matches (it indeed generated 2.3% fewer exact matches than the 5th participant) but often generated predicted segments closely corresponding to the true segments (e.g., it indeed generated 157.6% more "close matches" – i.e., accurate modulo a single token –

4 http://www.cnts.ua.ac.be/conll2003/ner/
5 http://www.cnts.ua.ac.be/conll2002/ner/
6 Erik Tjong Kim Sang, Personal communication, 25 Feb 2010.

Table 2
Rankings of the CoNLL'03 (English and German) and CoNLL'02 (Spanish) Shared Tasks participants according to the segment (Seg), token (T), and token & separator (TS) event spaces and to measures F_1^μ and F_1^M. The value in each cell represents the original rank the system obtained in the CoNLL'03 / CoNLL'02 evaluations, which use a segment-based F_1^μ exact-match model (1st, 4th, and 7th rows).

ENGLISH	Seg-F_1^μ	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
	T-F_1^M	1	2	3	4	11	6	8	5	7	10	9	14	15	12	13	16
		.885	.880	.865	.863	.857	.855	.853	.848	.847	.846	.833	.824	.822	.821	.815	.699
GERMAN	Seg F_1^μ	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
	T-F_1^M	1	9	3	2	4	7	6	5	8	11	10	13	12	14	15	16
		.726	.714	.713	.709	.701	.699	.697	.685	.680	.671	.652	.647	.644	.621	.582	.478
SPANISH	Seg F_1^μ	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				
	T-F_1^M	1	2	4	5	6	7	10	9	8	3	12	11				
		.823	.804	.775	.752	.752	.749	.741	.737	.732	.721	.681	.648				

Table 3
Spearman's rank correlation (averaged across the English, German, and Spanish tasks) $R(\eta', \eta'')$ between the results produced by the three evaluation models discussed in this section.

	Seg F_1^μ	TS-F_1^M	T-F_1^M
Seg F_1^μ	1.0	.832	.832
TS-F_1^M	.832	1.0	.990
T-F_1^M	.832	.990	1.0

than the 5th participant, and totally missed 6.9% fewer segments than the 5th participant). Conversely, our evaluation method demotes by 3 places each the participants that originally placed 5th, 7th and 12th. Several other participants are promoted or demoted by 2 places. The results are not much different in the 2003 German and 2002 Spanish data. In the 2003 German task we even have a participant that gains 8 positions (from 9th to 2nd place), while in the 2002 Spanish task one participant is downgraded from 3rd to 10th place (3rd from last!). These potentially very large differences clearly indicate that taking a clear stand between the two models is essential.

In order to check the level of correlation of the two models we have also computed (see Table 3) the Spearman's rank correlation $R(\eta', \eta'')$ between the rankings η' and η'' generated by the two models, where $R(\eta', \eta'') = 1 - \frac{6 \sum_{k=1}^{p} (\eta'(\hat{\Phi}_k) - \eta''(\hat{\Phi}_k))^2}{p(p^2-1)}$, p is the number of ranked participants, and $\eta(\hat{\Phi}_k)$ denotes the rank position of system $\hat{\Phi}_k$ in ranking η. We can see from Table 3 (whose values are obtained by averaging across the $R(\eta', \eta'')$ values obtained in the English, German and Spanish tasks) that the rankings produced

by the two models are fairly correlated ($R = .832$) but not *highly* so, confirming that taking a stand between the two is indeed important.

A potential criticism to using the set of all t-units as the event space (instead of, say, the set of all tokens) is that separators are given the same importance as tokens, which might seem excessive. For instance, the imperfect identification of the true segment "Barack Obama" via the predicted segment "Obama" results in one true positive and not one but *two* false negatives, which might be deemed too harsh a penalty. A potential solution to this problem consists in *weighting* tokens and separators differently, since F_1 can handle "weighted events" seamlessly. For instance, if we weigh separators half as much as tokens, a correctly tagged separator will count as "half a true positive"; accordingly, the example annotated sentence of Table 1 would obtain $F_1 = \frac{2*4.5}{2*4.5+1.5+0.5} = .818$. A similar solution could be adopted if different types of tokens are deemed to have different importance; for instance, heads might be weighted higher than modifiers in some applications, and last names might be weighted higher than first names when extracting person names.

Anyhow, in order to assess whether giving separators the same importance as tokens indeed constitutes a problem, we have re-evaluated the CoNLL'03 and CoNLL'02 results also according to a "token-only" F_1^M model (hereafter dubbed *T-F_1^M model*), i.e., a model which differs from our proposed model in that separators are not part of the event space, and are thus not the object of evaluation. The results are reported in the 3rd, 6th and 9th rows of Table 2. For the English data, we can see that the rankings are fairly similar, with only a few systems swapping places with the system next in the ranking (this happens for the systems placed 6th, 9th, and 14th in the TS-F_1^M ranking). For the German and Spanish data, the rankings are identical to the ones of the TS-F_1^M ranking. As a result, the Spearman's rank correlation R between the two rankings is very high ($R = .990$). All this indicates that the TS-F_1^M model does not place excessive emphasis on separators, which is good news.

Additionally, we should consider that separators tend to have even more negligible effects in IE tasks characterized by segments longer than the ones to be found in the CoNLL NER tasks. To see this, assume that, given a true segment σ containing n tokens, a tagger $\hat{\Phi}$ correctly recognizes only its subsegment containing the first $\frac{1}{2}n$ tokens. If $n = 2$, $\hat{\Phi}$ will obtain precision values of $\pi = \frac{1}{3}$ or $\pi = \frac{1}{2}$ (a very substantive difference) according to whether separators are considered or not in the evaluation. If $n = 100$, instead, $\hat{\Phi}$ will obtain precision values of $\pi = \frac{99}{199}$ or $\pi = \frac{50}{100}$, whose difference is almost negligible. Similar considerations hold for recall.

All in all, given that the difference between the rankings produced by the TS-F_1^M model and by the T-F_1^M model is small, and given that the former offers better theoretical guarantees than its token-only counterpart (since it guarantees that the correct identification of segment boundaries is properly rewarded), we think that the former should be preferred to the latter.

A scorer that evaluates a text annotated in the common IOB2 format according to *both* the segmentation F-score and the TS-F_1^M model can be downloaded at `http://www.isti.cnr.it/esuli/IOB2scorer.html`

6. Conclusion

We have argued that, in order to overcome the shortcomings of the standard "segmentation F-score" evaluation model for IE, the choices of event space and evaluation measure should be considered as two separate issues. For the former, we have proposed

using as the event space the set of all tokens and separators. We have shown that this (i) allows a correct appreciation of the degree of overlap between predicted and true segments, (ii) allows a fair evaluation of the ability of a system to correctly identify segment boundaries, (iii) has the consequence that the notion of a "true negative" is clearly defined, and (iv) allows the comparative evaluation of different IE systems to be carried out on the same contingency table. We have also argued that "tag switches" do not pose evaluation problems once different evaluations are carried out independently for different tags and then averaged. As for the evaluation measure, we have argued that, although there is nothing wrong with sticking to the standard F_1 measure, its macroaveraged version (F_1^M) is somehow more desirable, since it rewards systems that perform well across the entire tagset.

Finally, we should note that the notion of IE we have defined also allows a given t-unit to belong to more than one segment *for the same tag* c_i (we might thus dub this *multi-instance IE*). While this situation never occurs in simple applications of IE such as NER, there exist instances of IE in which this is the case. For example, in the tagset for opinion extraction defined in (Wiebe, Wilson, and Cardie 2005), it does happen that the same t-unit may belong to several segments for the same tag; e.g., in sentence "John wrote me that Mary said I love pizza", the segment "I love pizza" belongs to *two* overlapping segments of the INSIDE tag. Both the segmentation F-score and the evaluation model we have presented in this paper can only handle the single-instance IE case; we leave the issue of how to best evaluate multi-instance IE to further research.

References

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

De Sitter, An and Walter Daelemans. 2003. Information extraction via double classification. In *Proceedings of the ECML/PKDD'03 Workshop on Adaptive Text Extraction and Mining*, pages 66–73, Cavtat-Dubrovnik, KR.

Fawcett, Tom. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861—874.

Freitag, Dayne. 1997. Using grammatical inference to improve precision in information extraction. In *Proceedings of the ICML'97 Workshop on Automata Induction, Grammatical Inference, and Language Acquisition*, Nashville, US.

Freitag, Dayne. 2000. Machine learning for information extraction in informal domains. *Machine Learning*, 39:169–202.

Freitag, Dayne and Nicholas Kushmerick. 2000. Boosted wrapper induction. In *Proceedings of the 17th Conference of the American Association for Artificial Intelligence (AAAI'00)*, pages 577–583, Austin, US.

Joachims, Thorsten. 2005. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning (ICML'05)*, pages 377–384, Bonn, DE.

Krishnan, Vijay and Christopher D. Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL'06)*, pages 1121—1128, Sydney, AU.

Lavelli, Alberto, Mary Elaine Califf, Fabio Ciravegna, Dayne Freitag, Claudio Giuliano, Nicholas Kushmerick, Lorenza Romano, and Neil Ireson. 2008. Evaluation of machine learning-based information extraction algorithms: Criticisms and recommendations. *Language Resources and Evaluation*, 42(4):361–393.

Lewis, David D. 1995. Evaluating and optmizing autonomous text classification systems. In *Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR'95)*, pages 246–254, Seattle, US.

Sarawagi, Sunita. 2008. Information extraction. *Foundations and Trends in Databases*, 1(3):261—377.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Suzuki, Jun, Erik McDermott, and Hideki Isozaki. 2006. Training conditional random fields with multivariate evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (ACL/COLING'06)*, pages 217—224, Sydney, AU.

Tjong Kim Sang, Erik F. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning (CONLL'02)*, pages 155–158, Taipei, TW.

Tjong Kim Sang, Erik F. and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning (CONLL'03)*, pages 142–147, Edmonton, CA.

Tsai, Richard Tzong-Han, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung, and Wen-Lian Hsu. 2006. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7(92).

van Rijsbergen, Cornelis J. 1974. Foundations of evaluation. *Journal of Documentation*, 30(4):365–373.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):165–210.