Refined Tagging of Complex Verbal Phrases for the Italian Language

Simone Faro and Arianna Pavone

Dipartimento di Matematica e Informatica, Università di Catania,
Viale A. Doria 6, I-95125 Catania, Italy

Dipartimento di Scienze Umanistiche, Università di Catania
Piazza Dante 32, I-95124 Catania, Italy

The Prague Stringology Conference 2015
Prague 24-26 August, 2015

Recognition of terms and phrases which compose a text

A Verb Phrase is a syntactic unit consisting of one verbal form, combined with any other element, representing the verbal part of the speech.
The verb phrase is the central element in a sentence.

Recognition of terms and phrases which compose a text

- automatic information extraction from natural language texts
- semantic analysis of natural language texts
- automatic paraphrase
- knowledge bases construction
- automatic spelling
- part of speech tagging

Process of automatic language generation

- Easy problem
- Prearranged details for generation

Process of recognition, analysis and paraphrase

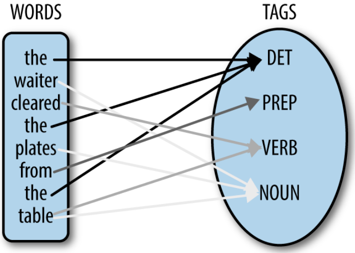
- Hard problem
- Presence of a large number of variants, concerning the syntax and the grammar
- Need for appropriate syntactic and semantic features

Part of Speech Tagging

The analysis of the parts of speech (PoS Tagging problem), with reference to the English language, is considered a simple problem today.

- The experimental results show that the PoS tagging solutions available for the English language can reach an accuracy up to 97%.

Such problem consists in analyzing a natural language text and in associating each part of the speech to a tag, selected from a predetermined set of tags. Such tag set could be more or less refined.



Part of Speech Tagging

The analysis of the parts of speech (PoS Tagging problem), with reference to the English language, is considered a simple problem today.

- The experimental results show that the PoS tagging solutions available for the English language can reach an accuracy up to 97%.

Such problem consists in analyzing a natural language text and in associating each part of the speech to a tag, selected from a predetermined set of tags. Such tag set could be more or less refined.

Applications

- tools for grammatical spell-correction of texts
- word processors
- e-mail clients
- electronic dictionaries
- search engines.

Part of Speech Tagging

PoS Tagging solutions are able to recognize the parts of speech by associating the terms in the text with the entries in some lexical Knowledge Base (KB), as:

- WordNet
- Multi-WordNet
- Euro-WordNet
- BabelNet

Lemmas in the KB include nouns, verbs, adjectives, adverbs etc. Each lemma or phrasal term in a KB, is associated to its sense, usually identified with one of the synsets related to the given term.

Recognition of Compound Terms

The compound phrases are difficult to be accurately recognized for three main reasons:

a) the terms which compose a compound phrase are themselves voices of the KB: *essere caduto* (*to have fallen*, past infinitive)

b) the terms composing a compound phrase may not appear contiguously in the text: *essere improvvisamente caduto* (*to have suddenly fallen*)

c) the conjugation of the terms contained in a compound verbal phrase may lead to a difficult recognition: *esserle caduta addosso* (*to have fallen on top of her*)

Verbal Tag Sets

- The reference tag set in PoS Tagging for the English language : Penn Treebank tag set (36 categories)
- The reference tag set in PoS Tagging for the Itlaian language:
  - EvalIta 2007: Treebank tag-set
    32 lexical categories , 6 verbal categories
  - EvalIta 2009: TANL tag-set
    37 elements with different morphological variants allowing the identification of 336 different elements.

Tag	Description	Examples (Italian)
VB	verb, lemma	*leggere, conoscere, andare*
VBD	verb, past	*leggevo, conobbi, andasti*
VBG	verb, gerund or present participle	*leggendo, conoscente, andando*
VBN	verb, past participle	*letto, conosciuta, andati*
VBP	verb, present, non-third singular person	*leggevamo, conosco, vai*
VBZ	verb, present, third singular person	*legge, conosce, va*

Tabella: The Treebank tag-set relative to verb phases.

Tag	**Description**	**Examples**
V	verb	*leggere, conosco, andato*
VA	auxiliary verb	*sono, eravamo, hanno*
VM	modal verb	*volevo, posso, dobbiamo*
Suffix	**Description**	**Examples**
-m	masculine	*letto, conosciuti, andato*
–f	feminine	*lette, conosciuta, andata*
–n	not specified	*leggo, conoscere, vanno*
–s	singular	*letto, conosci, va*
–p	plural	*lette, conoscevano, vanno*
–n	not specified	*leggere, conoscere, andare*
–1	first person	*leggevo, conosco, andammo*
–2	second person	*leggi, conoscevi, andrete*
–3	third person	*legge, conobbe, vanno*
–i	indicative	*leggo, conosceva, andavamo*
–m	imperative	*leggi, conosca, andate*
–c	subjective	*legga, conoscano, andassimo*
–d	conditional	*leggerei, conoscerebbe, andresti*
–g	gerund	*leggendo, conoscendo, andando*
–f	infinitive	*leggere, conoscere, andare*
–p	participle	*letto, conosciuta, andato*
–p	present	*leggo, conosco, vai*
–i	present perfect	*leggevo, conoscevi,*
–s	past	*lessi, conoscesti, andarono*
–f	future	*leggerà, conoscerete, andranno*
–c	clitics	*leggendocele, conoscilo*

Tabella: The TANL tag-set relative to verb phrases.

Italian Verb Phrases

In Italian, as in other languages, the verb phrase is the variable part of the speech and indicates an action, a state or a becoming in relation to a subject, expressed or implied, that does or undergoes an action. Some examples of verb phrases recognized by our tool are:

mangio	*(I eat)*
sono andato	*(I went)*
mi fu concesso	*(I was allowed)*
le è stato mandato	*(it was sent to her)*
mi pettino	*(I comb my hair)*

Main Tags: Verb Forms

Forms	Description	Examples
VSA	standard active	*capisco*
VSP	standard passive	*sono capito*
VPA	pronominal active	*avendolo capito*
VPP	pronominal passive	*avendomi capito*
VPR	pronominal reflexive	*essendomi capito*

Suffixes: Verb Values

Values	Description	Examples
:TR	transitive	*capissi*
:IN	intransitive	*andassi*

Suffixes: Verb Moods and Tenses

Tenses	Description	Examples
:IND	indicative	*avevo capito*
:CNG	subjective	*avessi capito*
:CND	conditional	*avrei capito*
:IMP	imperative	*capisci*
:GER	gerund	*avendo capito*
:PAR	participle	*capente*
:INF	infinitive	*capire*
Moods	**Description**	**Examples**
:PRE	present	*capisco*
:PAS	past	*capivo*
:FUT	future	*capirÚ*
:IMP	present perfect	*avevo capito*
:PRM	past perfect	*ebbi capito*
:TRA	distant past perfect	*avessi capito*
:FAN	future perfect	*avrÚ capito*

Suffixes: Gender, Number and Person

Gender	Description	Examples
:M	male	*Ë stato capito*
:F	female	*Ë stata capita*
:N	neuter	*abbiamo capito*
Number	**Description**	**Examples**
:S	singular	*capisci*
:P	plural	*capiamo*
:I	invariable	*capire*
Person	**Description**	**Examples**
:P0	impersonal	*aver capito*
:P1	first person	*abbiamo capito*
:P2	second person	*avete capito*
:P3	third person	*hanno capito*

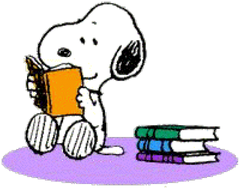
Pronominal Verb Forms

In Italian there are particular verbal forms with particles, called *clitics*. These clitics attach themselves to a word and they form a single unit.

leggerla (legger-la, to read it),
leggerne (legger-ne, to read some of them)
leggerci (legger-ci, to read to us).

Some of these verbs incorporate two clitics together, in these cases they are bi-pronominal verbs.

leggersela (legger-se-la, to read it to himself),
leggersene (legger-se-ne, to read some of them to himself)
leggerceli (legger-ce-li, to read them to ourselves).



Verb forms including a direct object.
They are built with the particles *-mi -ti -lo -la -li -le -ci* and *-vi*, where the particle assumes the function of direct object (with the meaning, respectively, of *me*, *you*, *him her*, *us*, *you* and *them*).
If the particles *-lo -la -li -le* are prefixed to the verb beginning with a vowel, the elision of the vowel is common: thus *l'amo* is equivalent to *la amo* (*I love her*).
Other examples are:

1. lo porti	*(you bring it)*
2. portarmi	*(to bring me)*
3. se l'avessi portata	*(if you had brought it)*

Verb forms including an indirect object.
Some pronominal forms use the particles *-mi* and its conjugations in gender and number, *-ti -gli -le -ci -vi*. In this case the pronominal particle is used as an indirect object (with the meaning of *to me*, *to you*, *to him*, *to her*, etc). This form is used with both transitive and intransitive verbs. Other examples are:

1. gli porti	*(you bring to him)*
2. portarmi	*(to bring to me)*
3. le avessi portata	*(you had brought to her)*

Verb forms including an adverb of place.

They are built by using the pronominal particle *-ci* or *-ne*, which have the function of adverb of place. The particle *-ci* is used with the meaning of *in that/this place* while the particle *-ci* is used with the meaning of *from that/this place*. In this context, the verb phrase *andarci* (*to go there*) can be paraphrased as *andare in quel luogo* (*to go in that place*). Other examples are:

1. arrivarci	*(to reach that place)*
2. ne vengo ora	*(I came now from there)*
3. lui ci viene	*(he came here)*

Verb forms including a partitive complement.
The particle *-ne* can be used also with the meaning *of that/this/them* with a partitive function. It can be applied to transitive and to intransitive verbs as well. Example of these verb phrases are:

1.	*parlarne*	*(to speak about that)*
2.	*ne avevamo spesi*	*(we spent some of them)*
3.	*ne porterÚ due*	*(I will bring two of them)*

Suffixes: Set of Clitics

Clitic	Description	Examples
:COC	object complement	*avermi portato*
:CTC	term complement	*avergli portata*
:CPC	place complement	*averci portati*
:CPF	partitive complement	*averne portate*

Ambiguity in the Recognition of Compound Terms

The compound tenses consist in (at least) two terms: an auxiliary verb, *essere* (*to be*) or *avere* (*to have*), conjugates in a simple tense, and a main verb conjugated in the past participle.

ho scelto
sono andato

In this context the past participle can be composed depending on the number or on the gender. The correct recognition (and the consequent tagging) of this verbal form creates some problems since in the Italian language the compound verbs can be composed in different ways.

Ambiguity in the Recognition of Compound Terms

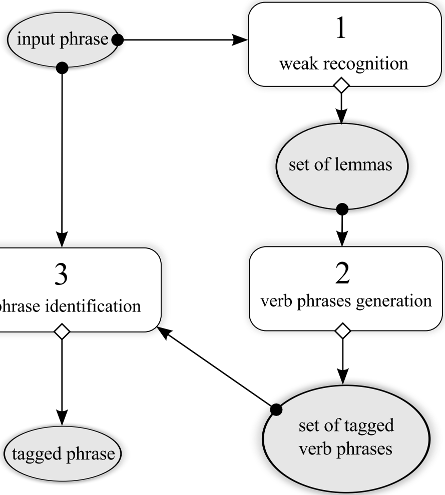
I chose the best solutions:
a1. ho scelto le migliori soluzioni
a2. ho scelte le migliori soluzioni

He has cheated us:
b1. ci ha ingannato
b2. ci ha ingannati

It was a news:
c1. lo è stato una novità
c2. lo è stata una novità

since we set ourselves that goal:
d1. essendocelo prefissati
d2. essendocelo prefissato

The Recognition Process



1. Weak Recognition Step

- The input text is tokenized and each term is associated with a (possibly empty) set of verb lemmas
- each term x_i is decomposed in two substrings p_i (a prefix) and s_i (a suffix) such that $x_i = p_i.s_i$. Any possible decomposition of the type $x_i = p_i.s_i$ is taken into account, with $|p_i| > 0$ and $|s_i| > 0$.
- If we find a prefix p_i which is equal to the radix of a verb v in our KB then we investigate if the corresponding suffix s_i could be a desinence of v. In such a case the verb v is returned as a lemma of x_i.

1. Weak Recognition Step

input phrase: *ce lo avevano portato* (*they had brought it to us*)

1. ce	$\emptyset$
2. lo	$\emptyset$
3. avevano	{*avere*}
4. portato	{*portare*}

2. Verb phrases generation step

- algorithm generates all possible verb phrases which are connected to the lemmas which have been identified at the previous step.
- let x_i a term of the input text t, and let $\{\ell_1, \ell_2, \dots, \ell_m\}$ the set of lemmas associated to x_i. The algorithm generates all possible verb phrases which are licensed by lemma ℓ_j, for $j = 1 \dots m$, by using a finite state model based on conjugation details stored in our .

2. Verb phrases generation step

Lemma *portare* (*to bring*):

portare → {	*porto*,	(VSA:TR:IND:PRE:N:S:P1)
	porti	(VSA:TR:IND:PRE:N:S:P2)
	porta	(VSA:TR:IND:PRE:N:S:P3)
	. . .	
	avessi portati	(VSA:TR:CNG:TRA:N:S:P2)
	avesse portati	(VSA:TR:CNG:TRA:N:S:P3)
	. . .	
	eravate state portate	(VSP:TR:IND:IMP:F:P:P2)
	erano state portate	(VSP:TR:IND:IMP:F:P:P3)
	. . .	
	ce lo avessi portato	(VSA:TR:CNG:TRA:N:S:P2:COC:CTC)
	ce lo avesse portato	(VSA:TR:CNG:TRA:N:S:P3:COC:CTC)
	. . . }	

3. Final Identification Step

During the final step of the process the algorithm identifies any possible verb phrase in the input text t by using information generated at the previous step.

- Let x_i be a term in t and let ℓ_j a lemma associated to x_i during the first step. Moreover let V_j be the set of all possible verb phrases which are licensed by lemma ℓ_j, generated at the previous step.
- The algorithm checks whenever each sequence $v \in V$ is equal to any subsequence of length k in t which involves the term x_i. This is done by comparing p with the subsequence $\langle x_h x_{h+1} \dots x_{h+k} \rangle$, for $h = \max(1, i-k) \dots \min(n, i+k)$.

Note:

Since each term can be involved in a single verb phrase, if two overlapping subsequences of t are recognized as verb phrases, then only the longest one is taken into account.

3. Final Identification Step

Sentence: *ce lo avevano portato* $(t = \langle x_1 \ldots x_5 \rangle)$

	verb phrase	lemma	tag
1.	*ce lo avevano*	*avere*	VSA:TR:IND:PAS:N:P:P3:CPC:COC
2.	*lo avevano*	*avere*	VSA:TR:IND:PAS:N:P:P3:COC
3.	*ce lo avevano portato*	*portare*	VSA:TR:IND:IMP:N:P:P3:CTC:COC
4.	*ce lo avevano portato*	*portare*	VSA:TR:IND:IMP:N:P:P3:CPC:COC
5.	*avevano*	*avere*	VSA:TR:IND:PAS:N:P:P3
6.	*avevano portato*	*portare*	VSA:TR:IND:IMP:N:P:P3
7.	*portato*	*portare*	VSA:TR:PAR:PAS:M:S:P1
8.	*portato*	*portare*	VSA:TR:PAR:PAS:M:S:P2
9.	*portato*	*portare*	VSA:TR:PAR:PAS:M:S:P3

Thank You!