

RTE

Recognising Textual Entailment

Valentina Bartalesi Lenzi

Matricola: 229572
C.d.L. Specialistica in Informatica Umanistica
Anno Accademico 2006/2007

Recognising Textual Entailment

Recognising Textual Entailment (RTE) è la capacità di un sistema di riconoscere, dati due frammenti di testo, se il significato dell'uno è inferito dall'altro.

Più correttamente: textual entailment è definito come una relazione direzionale tra due espressioni testuali (paio), denominate **Text (T)** e **Hypotesis (H)**.

Esempio:

T: "Il Primo Ministro Barham Salih è uscito illeso da un attentato terroristico"

H: "Il Ministro è stato il bersaglio di un attentato"

Il task risulta molto complesso data la presenza di fenomeni quali la variabilità e l'ambiguità del linguaggio.

Molte applicazioni di Natural Language Processing come **Question Answering**, **Information Extraction**, **Summarization** e **Machine Translation** necessitano di un modello che risolva il problema della variabilità e dell'ambiguità semantica.

Textual Entailment (1)

Secondo la logica deduttiva l'inferenza deriva dal significato principale delle relazioni che intercorrono tra due frasi.

In base a tale logica una frase ne inferisce un'altra **se e solo se** la veridicità dell'antecedente è garantita dalla veridicità della conseguente. Non è ammesso il contrario, la relazione di entailment è, infatti, una relazione unidirezionale.

È possibile caratterizzare l'entailment tra due frasi attraverso la loro combinazione usando il simbolo \rightarrow (material implication)

Entailment: $S_1 \vdash S_2$ iff $S_1' \rightarrow S_2'$ è una tautologia

Textual Entailment (2)

Si può vedere come viene applicata la definizione precedente attraverso un esempio di entailment tra frasi, che risulta da una regola di inferenza chiamata **conjunction entailment**.

La regola riconosce che ciascuna parte di una formula vera è anch'essa vera.

La frase (1.a), inferisce (1.b) e (1.c) se e solo se (2.a) e (2.b) sono tautologie.

Esempio:

- (1) a. Chester ate the cake and was sick.
b. Chester ate the cake.
c. Chester was sick.

- (2) a. $(\exists x[\forall y [\text{cake}'(y) \leftrightarrow x=y] \ \& \ (\text{eat}'(x)(\text{chester})] \ \& \ \text{sick}'(\text{chester}))$
 $\rightarrow \exists x[\forall y [\text{cake}'(y) \leftrightarrow x=y] \ \& \ (\text{eat}'(x)(\text{chester}))]$.
b. $(\exists x[\forall y [\text{cake}'(y) \leftrightarrow x=y] \ \& \ (\text{eat}'(x)(\text{chester})] \ \& \ \text{sick}'(\text{chester}))$
 $\rightarrow \text{sick}'(\text{chester})$.

Textual Entailment (3)

Si può dimostrare che entrambe le formule in (2) sono tautologie usando il **metodo della tabella delle verità**.

p: Chester ate the cake

q: Chester was sick

p & q: la congiunzione tra p e q

$(p \& q) \rightarrow p$ questa è una tautologia poichè l'implicazione finale è sempre vera, non importa quali valori siano assegnati a p e q.

p	q	p & q	$(p \& q) \rightarrow p$
t	t	t	t
t	f	f	t
f	t	f	t
f	f	f	t

PASCAL network e RTE Challenge

Le competizioni RTE sono organizzate nell'ambito della rete di eccellenza PASCAL.

L'obiettivo di PASCAL è quello di costituire un istituto europeo all'avanguardia nello studio dei metodi principali di:

- pattern analysis
- statistical modelling
- computational learning

La rete di eccellenza PASCAL è coordinata da John Shawe-Taylor (Coordinatore scientifico) dell'University College di Londra e Steve Gunn (Coordinatore operativo) dell'University of Southampton.

La rete di eccellenza PASCAL si interessa, in particolare, di:

- Computational, Information-Theoretic Learning with Statistics
- Learning/Statistics & Optimisation
- Theory & Algorithms
- Machine Vision
- Speech
- Haptics
- Brain Computer Interfaces
- Natural Language Processing
- Information Retrieval & Textual Information Access
- Multimodal Integration
- User Modelling for Computer Human Interaction

RTE Challenge (1)

1) **Creazione di una collezione** di coppie (paia) di piccoli frammenti di testo, rispettivamente *Text* e *Hypothesis*.

2) **Esercizio**: ai sistemi partecipanti viene chiesto di decidere per ciascun paio di T/H se T inferisce H, assegnando ad ogni coppia il valore TRUE o FALSE.

3) **Valutazione**: i risultati dei sistemi vengono comparati con un gold standard creato manualmente.

La competizione comprende diversi sotto-task, ciascuno per una diversa tecnologia di riferimento: **Information Retrieval (IR)**, **Question/Answering (Q/A)**, **Information Extraction (IE)**, **Summarization (SUM)**.

RTE Challenge (2)

Evoluzione del task di RTE:

1) 15 giugno 2004 - 10 aprile 2005 à **prima campagna di valutazione** organizzata dalla Bar Ilan University (Israele) e ITC-irst (Trento): 17 gruppi di ricerca partecipanti.

2) 1 ottobre 2005 - 10 aprile 2006 à **seconda campagna di valutazione**. Organizzata Bar-Ilan University (Israele), CELCT (Trento), Microsoft Research (USA) e MITRE (USA): 23 gruppi partecipanti.

3) 20 dicembre 2006 - 28 giugno 2007 **terza campagna di valutazione** organizzata da Bar-Ilan University (Israele), CELCT (Trento), ITC-irst (Trento) Microsoft Research (USA), MITRE (USA), USC/ISI (USA): 26 gruppi partecipanti

Linee guida per la creazione della collezione (1)

1. KEEP IT SIMPLE

La collezione deve essere compresa dai computer e non dagli umani. Ciò che appare di facile comprensione per l'umano può risultare estremamente difficile per un computer. Per esempio non è facile per un computer distinguere tra "lanciare un attacco" da "attaccato".

Quindi:

- § Il 75% delle T nella collezione deve essere costituito da brevi frasi e solo il 25% da frasi più lunghe (uno o due paragrafi). Le H devono essere costituite **solo** da frasi brevi.
- § Devono essere **evitate** frasi o paragrafi con **strutture complesse**.
- § Il linguaggio usato deve essere quello **contemporaneo**: l'inglese semplice. Non sono ammessi termini arcaici o dialettali.

Linee guida per la creazione della collezione (2)

2. WORLD KNOWLEDGE

Le inferenze che contribuiscono all'entailment sono spesso basate solo su ciò che è riportato nel testo. Viene comunque ammesso un **certo grado di conoscenza del mondo**.

Limiti: perché un computer riesca a risolvere l'entailment del seguente paio, sarebbe necessario un grado di conoscenza del mondo eccessivamente elevato (la macchina dovrebbe riuscire a dedurre che in inverno nevica)

T: "Stava nevicando a Gerusalemme quando è nato il bambino"

H: "Il bambino è nato in inverno"

Questa coppia di T/H, e quelle che richiedono un world knowledge simile, non dovrebbero essere inserite nella collezione.

Linee guida per la creazione della collezione (3)

3. ENTAILMENT DIFFICULTY

Variabilità linguistica

L'entailment contempla una certa variabilità linguistica, è comunque richiesto che il numero di questi fenomeni all'interno della collezione sia il più possibile ridotto.

Esempio

T: "Il partito ha eletto la signora Gandhi come guida in parlamento"
H: "Sonia Gandhi guida il partito in parlamento"

T mostra tre tipi di variazione rispetto a H:

- § lessicale: titoli (signora Gandhi à Sonia Gandhi)
- § morfologica: nominalizzazione (guida à guidare)
- § sintattica: cambiamento della struttura della frase (Sonia Gandhi à complemento oggetto; signora Gandhi à soggetto)

Linee guida per la creazione della collezione (4)

Tempi verbali

Le paia di T/H devono essere modificate al fine di coordinare i loro tempi verbali evitando, in questo modo, casi controversi di entailment.

Anafora

T potrebbe contenere un'anafora e il termine che indica l'entità di riferimento potrebbe non essere menzionato nella stessa frase ma in una precedente. In questo caso, l'anafora deve essere sostituita con il termine che indica l'entità corrispondente.

Esempio

T: "Orhan Pamuk è un grande scrittore turco. È lui che ha vinto il premio nobel per la letteratura quest'anno"

Se T è costituita solo dalla seconda frase allora deve essere modificata in questo modo:

T: "È Orhan Pamuk che ha vinto il premio nobel per la letteratura quest'anno"

Linee guida per la creazione della collezione (5)

4. QUANTITIES

Devono essere inseriti nella collezione 200 esempi per ciascun sotto-task. Devono essere selezionati esempi differenti in termini di:

- § vocabolario
- § lunghezza
- § livello di difficoltà
- § esempi positivi e negativi

Il numero di paia positive e negative per ciascuna Hypotesis non è necessario che sia bilanciato.

Linee guida per il task di IR (1)

Documenti web, come i risultati di una query forniti da un motore di ricerca, devono essere usati per la generazione delle T. Possono essere usati differenti motori di ricerca: Altavista, Google, Yahoo, Microsoft.

Le T devono essere estratte da articoli di giornale e devono presentare un linguaggio e una sintassi comuni al Web, con costrutti non artificiali.

Le H devono essere utilizzate come query nel motore di ricerca. Dovrebbero essere concepite in modo tale che il motore trovi il maggior numero possibile di risultati.

Esempio:

H: "La paura per gli spazi pubblici (agorafobia) è un disturbo diffuso"

Può essere modificata in:

H: "L' agorafobia è un disturbo diffuso"

Linee guida per il task di IR (2)

Esempi

TEXT	HYPOTHESIS	JUDGEMENT
Agoraphobia is a widespread disorder.	The fear of open or public places (Agoraphobia) is a widespread disorder.	true
Studies have shown that agoraphobia, literally "fear of the marketplace," is often coupled with panic disorder	The fear of open or public places (Agoraphobia) is a widespread disorder.	false

Linee guida per il task di Q/A (1)

Le paia per il task di Q/A devono essere generate seguendo le seguenti procedure:

- § Modificando le risposte restituite da sistemi partecipanti alla campagna di valutazione di Q/A (CLEF)
- § Revisionando le paia riprese da l'esercizio AVE (Answer Validation Exercise) e non utilizzate nella campagna di valutazione di Q/A
- § Recuperando risposte manualmente utilizzando sistemi di Q/A e documenti pubblicati sul web
- § Revisionando l'output di 3 sistemi di Q/A:
 1. Il sistema del gruppo di Dan Moldovan
 2. Il sistema del gruppo di Sanda Harabagiu
 3. Il sistema di Zhiping Zheng

Esempio:

Q: "Quante sono state le vittime della guerra in Ruanda?"

A/T: "In Ruanda le vittime sono state in media 8000 al giorno per circa 100 giorni"

H: "In Ruanda la guerra ha provocato circa 80000 vittime"

Linee guida per il task di Q/A (2)

Esempi

TEXT	HYPOTESIS	JUDGEMENT
Hebron, located south of Jerusalem in the Judean hills, is home to approximately 120,000 Arabs, 500 Jews, and three Christians.	Hebron is located in the Judean hills.	true
The Hebron District is located in the southernmost portion of the West Bank, 36 km south of Jerusalem City.	Hebron is located in Jerusalem City.	false

Linee guida per il task di IE (1)

IE è un processo attraverso il quale specifici tipi di informazioni (es. date, luoghi) sono estratti da specifici testi seguendo un tema preciso (es. attacchi terroristici).

Le informazioni vengono poi riformulate in linguaggio naturale e sono usate come H.

Le paia T/H vengono generate attraverso differenti approcci:

1. Dall'output del task ACE-2005 RDR (Relation Detection and Recognition)
2. Attraverso l'output di sistemi di IE utilizzati sul dataset di MUC-4 TST3
3. Alcune paia vengono create manualmente sulla base del dataset di MUC-4 TST 3 e degli articoli collezionati per il task ACE-2005 RDR

Linee guida per il task di IE (2)

Esempi

TEXT	HYPOTESIS	JUDGMENT
The Mall School Richmond has replaced its aging Windows 2000 PCs with a state of the art Linux thin-client network, offering access to a wealth of educational software at a fraction of the cost of an upgraded Microsoft-based solution.	The Mall School is a user of Linux.	true
The Mall School in Richmond has replaced its aging Windows 2000 PCs with a state of the art Linux thin-client network, offering access to a wealth of educational software at a fraction of the cost of an upgraded Microsoft-based solution.	The Mall School is a customer of Microsoft.	false

Linee guida per il task di Summarization (1)

Le paia per questo task vengono collezionate utilizzando l'output di tre sistemi di summarization:

1. Newsblaster (Columbia University)
2. NewsInEssence (Dragomir R. Radev e Jahna Otterbacher, University of Michigan)
3. Pyramids (Columbia University)

PYRAMIDS

Le T sono costituite dall'output dei sistemi

Le H dalle SCU (Summarization Content Unit)

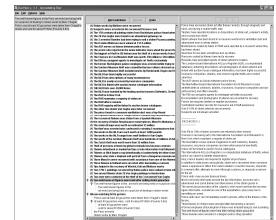


Fig.1 Screenshot dell'interfaccia di Pyramids

Linee guida per il task di Summarization (2)

Esempi

TEXT	HYPOTESIS	JUDGMENT
Rubens painting Aurora, stolen from a Spanish museum six years ago was recovered in Florida	The recovered Rubens was stolen from a Spanish Museum	true
Amsterdam police said Wednesday that they have recovered stolen lithographs by the late U.S. pop artist Andy Warhol worth more than \$1 million	Millions of dollars of art were recovered, including works by Dali	false

Linee guida per l'annotazione (1)

Giudizio sull'entailment:

1. **L'ipotesi deve essere inferita per intero dal testo.** Il giudizio dovrebbe essere ritenuto falso se H include porzioni di testo che non possono essere inferite da T.
2. **Devono essere ignorati i tempi verbali.** Per esempio viene accettata come true l'ipotesi *Yahoo ha acquistato Overture* anche se nel testo Yahoo ha solo annunciato l'imminente acquisto.
3. **Deve essere ammesso un certo grado di conoscenza del mondo.** Per esempio, un'azienda ha un AD, l'amministratore è un dipendente, il dipendente è una persona.
4. **È considerato corretto un entailment che è solo probabile ma non certo.** Per esempio *Maria ha molto criticato l'articolo* non è la stessa cosa di *Maria ha rifiutato l'articolo* ma è comunque molto probabile che questo accada e quindi l'entailment è considerato true.
5. **L'entailment è una relazione unidirezionale.** L'ipotesi deve essere inferita completamente dal testo ma il testo non deve essere inferito dall'ipotesi.

Linee guida per l'annotazione (2)

Entailment *strict* e *plausible*

L'entailment viene definito *strict* se la verità espressa nell'ipotesi è chiaramente ribadita dal testo.

L'entailment viene definito *plausible* se dato un testo, la verità espressa nell'ipotesi è altamente probabile ma non certa.

Esempi:

Entailment *strict*

T: "L'Italia e la Germania hanno giocato già due volta ma non hanno battuto nessuno"

H: "Né l'Italia né la Germania hanno ancora vinto"

Entailment *plausible*

T: "Sonia Gandhi potrebbe essere sconfitta nelle prossime elezioni in India se tra adesso e il 2009 il livello di disoccupazione non scenderà"

H: "Le prossime elezioni in India saranno nel 2009"

Esempio di collezione di paia

Le paia vengono annotate contemporaneamente da 3 annotatori diversi.

Perché l'entailment venga considerato positivo tutti e tre gli annotatori devo aver espresso come giudizio "YES".

Per l'annotazione non esiste un apposito tool ma viene utilizzato un semplice foglio Excel

Esempio di file cross annotato: [test_set_cross_annotation.xls](#)

Ogni annotatore, poi, in un apposito foglio excel per ogni paia, oltre a stabilire se esiste o meno l'entailment, riporta il proprio giudizio su:

§ la necessità di avere world knowledge per poter valutare l'entailment

§ se l'entailment è strict o plausible

§ il grado di difficoltà (regular, difficult) nello stabilire l'entailment

Esempio di file con giudizi sull'entailment: [test_set_ann3.xls](#)

Modifiche nel terzo RTE Challenge

Le linee guida per la creazione della collezione e l'annotazione sono rimaste identiche per i primi due anni.

In occasione del **terzo Recognising Textual Entailment Challenge** sono state apportate le seguenti modifiche:

- 1) Adozione di un limitato numero di T più estese, in modo da fornire una maggiore porzione di contesto facilitando il funzionamento di sistemi di analisi del discorso;
- 2) Creazione di un *RTE Resource Pool* dove i partecipanti possono condividere le risorse che usano;
- 3) Utilizzo dell'output di Pyramids per la creazione della collezione per il task di Summarization;
- 4) Introduzione di un nuovo task "Extending the Evaluation of Inferences from Texts" diviso in due sotto task: nel primo viene richiesto al sistema di distinguere il giudizio negativo sull'entailment tra "NO" e "UNKNOWN", nel secondo viene valutata la capacità del sistema di giustificare la risposta data.

Misure di valutazione

Valutazione principale: **classificazione**

§ Comparazione del giudizio sull'entailment

§ Accuracy

$$\text{accuracy} = \frac{\text{number of True Positives} + \text{number of True Negatives}}{\text{numbers of True Positives} + \text{False Positives} + \text{False Negatives} + \text{True Negatives}}$$

§ Baseline: 60%

Second RTE Challenge

- § 23 gruppi
- § 41 runs
- § 13 gruppi partecipanti per la prima volta

Paese	Numero di gruppi
USA	9
Italy	3
Spain	3
Netherlands	2
UK	2
Australia	1
Canada	1
Ireland	1
Germany	1

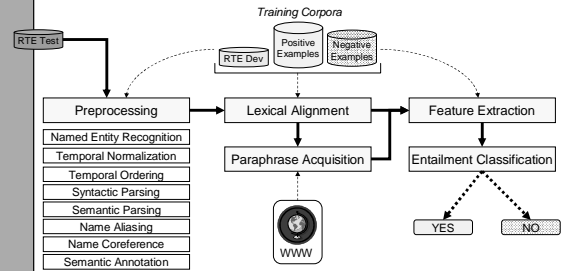
Risultati del Second RTE Challenge

Primo autore (Gruppo)	Accuracy
Hickl (LCC)	75,4%
Tatu (LCC)	73,8%
Zanzotto (Milano e Roma)	63,9%
Adams (Dallas)	62,6%
Bos (Roma e Leeds)	61,6%
11 gruppi	58,1%-60,5%
7 gruppi	52,9%-55,6%

Risultati divisi per task

	Average Accuracy	Risultati
SUM	67,9%	84,5%
IR	60,8%	74,5%
QA	58,2%	70,5%
IE	52,2%	73,0%
Totale	59,8%	75,4%

RTE con LCC's GROUNDHOG System



Preprocessing (1)

Groundhog inizia il processo di RTE annotando le paia con un insieme di informazioni lessico-sematiche:

Named Entity Recognition

CiceroLite NER software permette di annotare differenti tipi di Named Entity:

[The Bills]_{SPORTS_ORG} plan to give the starting job to [J.P. Losman]_{PERSON}

Coreferenza

CiceroLite permette anche di annotare la coreferenza

[The Bills]_{ID=01} now appear ready to hand the reins over to [one of [their]_{ID=01} two-top picks]_{ID=02} from a year ago in [quarterback]_{ID=02} [J.P. Losman]_{ID=02}, [who]_{ID=02} missed most of last season with a broken leg.

Preprocessing (2)

Riconoscimento e normalizzazione di Espressioni Temporal

TASER system viene usato per il riconoscimento e la normalizzazione di espressioni temporali

The Bills [now]_{2006/01/01} appear ready to hand the reins over to one of their two-top picks from [a year ago]_{2005/01/01-2005/12/31} in quarterback J.P. Losman, who missed most of [last season]_{2005/01/01-2005/12/31} with a broken leg.

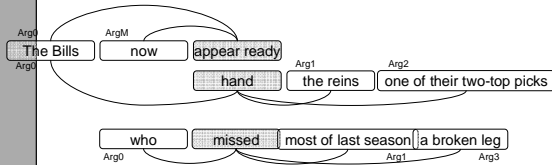
POS Tagging and Syntactic Parsing

Utilizzo del Brill POS tagger e del Collins Parser

Preprocessing (3)

Semantic Parsing

The Bills now **appear ready** to **hand** the reins over to one of their two-top picks from a year ago in quarterback J.P. Losman, who **missed** most of last season with a broken leg.



Preprocessing (4)

Annotazione semantica

Polarità: ai predicati viene assegnata polarità negativa o positiva.

Members of Iraq's Governing Council **refused** to [sign]_{FALSE} an interim constitution

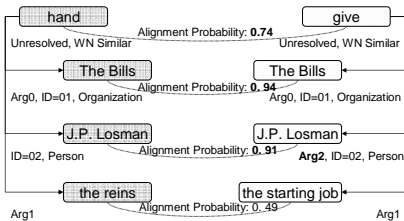
Verbi fattivi: presuppongono che la proposizione che introducono sia vera

Verbi non-fattivi: non presuppongono che la proposizione che introducono sia vera

Verbi "unresolved": verbi che indicano azioni verbali (*deny, claim*)
 verbi che indicano opinioni (*think, believe*)
 verbi che indicano incertezza (*be uncertain*)
 verbi che indicano intenzione (*scheme, plot, want*)
 verbi in contesti condizionali (*whether, if*)

Lexical Alignment

L'annotazione lessico-semantica viene utilizzata per cercare di identificare le corrispondenze tra chunk e token di T e H



Paraphrase Acquisition (1)

Groundhog identifica i sintagmi verbali per ogni paio di T e H. (Dolan et al. 2004, Barzilay and Lee 2003, Shinyama et al. 2002).

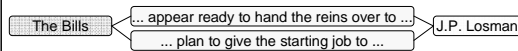
L'output dell'allineatore può essere utilizzato per identificare le "target region", cioè le porzioni di testo di T e H che hanno alta probabilità di essere parafrasi l'una dell'altra.

The Bills now **appear ready to hand the reins over** to one of their two-top picks from a year ago in quarterback **J.P. Losman**, who missed most of last season with a broken leg.

The Bills **plan to give the starting job to J.P. Losman**.

Paraphrase Acquisition (2)

Se si deduce che una porzione di testo di H è parafrasi di una porzione di testo di T si può giungere alla conclusione che le due porzioni di testo abbiano lo stesso significato.



Non tutte le frasi contenenti le due entità allineate sono vere parafrasi...

- ... may go with quarterback ...
- ...could decide to put their trust in...
- ..might turn the keys of the offense over to..

Entailment Classification

Utilizzo di un sistema di machine learning classifier per determinare quando esiste una relazione di entailment tra le paia.

Vengono utilizzate 4 feature per l'entailment classification:

1. **Alignment Features:** valutazione della lunghezza delle stringhe comuni per T e H. Maggiore è la lunghezza delle stringhe comuni, maggiore è il valore assegnato a quella coppia dal classificatore. Valuta, inoltre, i chunk e token corrispondenti in T e H allineati a livello lessico-semantic.
2. **Dependency Features:** valuta se sono state assegnate le stesse relazioni semantiche alle entità corrispondenti in T e H.
3. **Paraphrase Features:** valuta se ci sono porzioni di testo di H che possono essere parafrasi di T. Maggiore è il numero di parafrasi, maggiore è il valore assegnato a quel paio.
4. **Semantic Features:** valutazione della polarità dei predicati presenti in T e H e dei valori assegnati ai verbi. Maggiore è la coincidenza, maggiore è il valore assegnato alla coppia.

Problemi aperti

1) Il world knowledge e i dati di training sono più importanti dei metodi di inferenza testuale?

2) O piuttosto dato un maggior numero di dati di training e di world knowledge, le differenze tra i vari metodi di inferenza testuale diventeranno più significative?

Bibliografia

- S. Pinker "L'istinto del Linguaggio. Come la mente crea il linguaggio", Mondadori, 1998.
- I.Dagan, O.Glickman, B.Magnini, 2006. *The PASCAL Recognising Textual Entailment Challenge*. In Quinonero-Candela et al., editor, MLCW 2005, LNAI Volume 3944. Springer-Verlag.
- R.Bar-Haim, I.Dagan, B.Dolan, L.Ferro, D.Giampiccolo, B.Magnini and I.Szpektor, 2006. *The Second PASCAL Recognising Textual Entailment Challenge*. In Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venice, Italy.
- I.Dagan, O.Glikman, 2004. *Probabilistic textual entailment: Generic applied modeling of language variability*. PASCAL workshop on Text Understanding and Meaning.
- J.Burger, L.Ferro, 2005. *Generating an entailment corpus from news headlines*. In Proceedings of ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Ann Arbor, Michigan. Association for Computational Linguistics.
- A.Hickl, J.Bensley, J.Williams, K.Roberts, B.Rink, Y.Shi, 2006. *Recognizing Textual Entailment with LCC's GROUNDHOG System*. Proceedings of the Challenges Workshop, Venice. Kurimo-Creutz-Lagus.
- R. Cann "Formal semantics - An introduction", Cambridge University Press, 1993

Sitografia

PASCAL Network: <http://www.pascal-network.org/Challenges/RTE2>

Recognising Textual Entailment Challenges:
<http://www.pascal-network.org/Challenges/RTE/>

The Second PASCAL Recognising Textual Entailment Challenge (RTE-2):
<http://www.pascal-network.org/Challenges/RTE2/>

3rd PASCAL Textual Entailment Challenge and Resources Pool:
<http://www.pascal-network.org/Challenges/RTE3>

Textual Entailment Portal:
http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Portal

PASCAL Challenges Workshop 2:
http://seminars.ijs.si/pascal/2006/pcw2_venice/default.asp

RTE3 Optional Pilot Task: Extending the Evaluation of Inferences from Texts:
<http://nlp.stanford.edu/RTE3-pilot/>

Textual Entailment Resource Pool:
http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool

B. MacCartney, C. D. Manning "Natural Logic for Textual Inference"
<http://nlp.stanford.edu/~wcmac/papers/natlog-wtep07.pdf>

”