UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Valutare i sistemi per il Trattamento Automatico della Lingua: il task di Frame Labeling in Evalita 2011

Candidato: *Carmela Cinquesanti*

Relatore: *Prof. Alessandro Lenci*

Correlatore: *Prof.ssa Maria Simi*

Anno Accademico 2010-2011

Indice

1.Introduzione

La presente relazione documenta la realizzazione di un progetto basato sull'annotazione di frasi in lingua italiana mediante strutture semantiche in stile FrameNet. I risultati della prima parte del progetto sono stati impiegati come dati di test da sottoporre ai sistemi di Frame Labeling nella campagna di valutazione Evalita 2011, mentre i risultati della seconda parte sono andati ad arricchire la ISST (Italian Syntactic-Semantic Treebank) con nuove istanze di *frames* rintracciate. Il lavoro viene pertanto inquadrato in uno scenario metodologico più ampio, che si concentra sul tema della valutazione applicata ai sistemi per l'elaborazione del linguaggio naturale (dall'inglese Natural Language Processing, NLP).

Lo scopo dell'attività di valutazione è misurare le prestazioni di un algoritmo o di un sistema, al fine di determinare se, e in che misura, il sistema in questione raggiunge gli obiettivi dei suoi ideatori. In altri termini, si verifica se il prodotto risponde in maniera adeguata e coerente alle esigenze degli utenti per cui è stato creato. La ricerca nel campo della valutazione attira una notevole attenzione da parte degli sviluppatori dei sistemi, in quanto la definizione di precisi criteri valutativi permette di specificare in maniera altrettanto precisa i problemi da affrontare nei task di NLP. Inoltre, la presenza di un particolare insieme di criteri e metriche di valutazione comporta la possibilità per gli sviluppatori di confrontarsi su un determinato compito, comparando non solo le soluzioni adottate nelle diverse implementazioni, ma anche i risultati ottenuti. Il vantaggio di un simile approccio consiste nella realizzazione di un ambiente condiviso, in cui la competizione acquista un significato positivo e la spinta a ottimizzare i sistemi di elaborazione del linguaggio genera risultati sempre più validi. Anche il livello di complessità dei task risente dei benefici effetti della valutazione: difatti, la possibilità di definire compiti molto specifici permette di superare la vaghezza di semplici operazioni come la comprensione del testo o la generazione automatica del linguaggio.

Le campagne di valutazione sono delle iniziative che promuovono la valutazione dei sistemi di NLP. Una campagna di valutazione consiste in una competizione tra più squadre: dato uno specifico task, ciascuna squadra partecipa all'esecuzione del

compito con il proprio sistema. I risultati delle varie squadre vengono poi messi a confronto al fine di stabilire il sistema più efficiente in quel determinato task. La competizione si articola in due fasi fondamentali: nella prima fase gli organizzatori della campagna creano e distribuiscono dei dati di addestramento (training o development data) per lo sviluppo dei sistemi di NLP; nella seconda fase, i sistemi vengono valutati su dati di prova (test data) e comparati fra loro. I risultati delle elaborazioni sono le basi da cui partire per valutare pregi e difetti dei sistemi partecipanti, discutendo i punti salienti della competizione in una serie di *workshop*. Oltre a fornire nuove conoscenze tecnologiche, le campagne di valutazione creano uno stretto rapporto di collaborazione tra gli informatici e i linguisti; tale sinergia è sostanziale nella definizione dei repertori di dati comuni, nella scelta dei criteri di valutazione e nell'organizzazione dell'elaborazione dei dati.

La nascita delle campagne di valutazione risale al 1987, nel quadro dei progetti DARPA sviluppatisi negli Stati Uniti. In questo contesto, vennero organizzate le prime campagne di valutazione per strumenti di elaborazione del linguaggio e di comprensione testuale.

In Italia, nasce nel 2007 il progetto Evalita, che si occupa della valutazione di strumenti di elaborazione per la lingua italiana. Proprio nell'ambito della terza edizione di questa iniziativa è stata eseguita la valutazione di sistemi di annotazione semantica descritta nella relazione. Il resoconto presenta gli strumenti e le metodologie adottate nella valutazione, prima descrivendo i task di *Frame Labeling over Italian Texts* e *Super Sense Tagging*, poi concentrandosi sulla partecipazione al compito di Frame Labeling. In particolare, il lavoro di preparazione dei dati di test si è svolto secondo una metodologia semiautomatica, rivedendo a mano un campione di frasi pre-annotate automaticamente da Moses, uno strumento di Sematic Role Labeling dell'Università di Roma, Tor Vergata.

2. Le campagne di valutazione

Le campagne di valutazione costituiscono un aspetto fondamentale nell'ambito dell'elaborazione del linguaggio naturale, il campo disciplinare al confine tra la linguistica, l'informatica e l'intelligenza artificiale che si occupa dell'interazione tra i calcolatori elettronici e le informazioni fornite in linguaggio naturale. Le campagne di valutazione acquistano rilevanza a partire dall'interesse dei gruppi di consumatori (inclusi sviluppatori di sistemi, finanziatori e utenti finali) per i risultati della valutazione. Nello specifico, questo interesse dialoga con il ciclo di vita delle tecnologie più avanzate, al fine di determinare i tipi di valutazione appropriati ai differenti stadi di maturità di un sistema: lo studio si estende da valutazioni basate sulla componente tecnologica[1] nei primi stadi, ad analisi incentrate sull'interazione con l'utente[2], nelle fasi di maggiore maturità del sistema. Le domande che si pongono gli utenti finali riguardano gli obiettivi dell'elaborazione del linguaggio naturale, la velocità di sviluppo di questo campo di studi e l'utilità di specifiche implementazioni, e aumentano di pari passo con l'evoluzione delle tecnologie in prodotti commerciali.

La scelta di un metodo valutativo dipende dalla specifica fase di crescita in cui il sistema si trova: possono distinguersi quattro momenti fondamentali: la ricerca, il prototipo avanzato, il prototipo operazionale e infine il prodotto ultimato. L'oggetto e il metodo della valutazione variano a seconda dello stadio esaminato, e in ciascuno di questi si delinea un quadro del processo di sviluppo, evidenziando tanto i progressi ottenuti, quanto le aree da migliorare; nei primi momenti di vita del sistema, poiché la tecnologia è ancora fragile e i soggetti maggiormente coinvolti sono gli sviluppatori e i finanziatori, la valutazione avviene a livello del componente tecnologico. Lo stadio, ancora poco maturo, del prototipo operazionale offre alla tecnologia emergente la possibilità di supportare degli specifici studi sul campo. Soltanto in un momento successivo, è possibile effettuare esperimenti più precisi, focalizzati soprattutto sul *feedback* da parte degli utenti reali: è il caso del prodotto finito, che viene valutato in base all'approvazione e al rendimento sul mercato.

[1] La valutazione basata sulla tecnologia si concentra sul rendimento dei componenti tecnologici sottostanti, misurando p.es. velocità, precisione e prestazioni.

[2] La valutazione incentrata sull'utente considera l'interazione tra l'utente e il sistema.

Oltre che dagli stadi dell'evoluzione, il metodo valutativo dipende dalla natura del sistema, ossia dal tipo di input e output. Si possono distinguere tre classi di appartenenza:

1. Sistemi di *analisi*: l'input consiste in espressioni linguistiche, di cui viene fornita una rappresentazione astratta. È il caso dei sistemi di estrazione e recupero di informazioni, di *topic identification* e di analisi. La valutazione può avvenire in relazione a uno standard di riferimento, ossia un insieme di output corretti per un certo gruppo di input di prova. È possibile assegnare un punteggio all'output del sistema, per creare un confronto con lo standard di riferimento;
2. Sistemi che producono *output linguistici*: si tratta di sistemi per la traduzione, il riassunto e la generazione di testo. In questo caso, può non esistere un unico output corretto per un determinato input: a uno stesso testo, infatti, potrebbero corrispondere più traduzioni o riassunti. Le misure valutative predominanti sono la qualità e l'informatività;
3. Sistemi *interattivi*: l'utente e il computer interagiscono in uno scambio reciproco di informazioni per raggiungere un obiettivo. Questo è lo scenario di gran lunga più complesso, perché il successo del sistema dipende da entrambe le entità che prendono parte all'interazione.

2.1 Valutazione dei sistemi di analisi

Nell'insieme dei sistemi di analisi rientrano vari tipi di applicazioni, a seconda della rappresentazione astratta prodotta, e le misure che solitamente supportano la valutazione di questi sistemi sono la *recall*, la precisione e il tasso di errore. La recall è il rapporto tra il numero di istanze correttamente individuate e il numero totale di istanze che possono essere considerate rilevanti nel testo. Nel caso dell'Information Retrieval questa grandezza rappresenta la probabilità che un documento pertinente venga recuperato. La precisione misura il numero di istanze rilevanti tra tutte quelle individuate. Il tasso di errore somma la quantità di aggiunte, cancellazioni e sostituzioni, divisa per il numero totale di istanze pertinenti. Fanno parte dei sistemi di analisi:

1. la *segmentazione*, che comprende la segmentazione di parole, la segmentazione del discorso (per la lingua parlata) e del racconto, e l'individuazione dei confini di frase;
2. il *tagging* a vari livelli, inteso come individuazione ed etichettatura delle unità. Ne fanno parte il *Part of Speech tagging*, il tagging del discorso (che prevede dapprima la segmentazione del testo in unità elementari e poi la ricostruzione delle relazioni inter-proposizionali), il tagging del senso di parola e l'analisi morfologica;
3. l'*estrazione di informazione* che recupera automaticamente informazioni da documenti di testo o audio non strutturati e le organizza in schemi strutturati, come *frames* e tabelle;
4. *threading e ordinamento dei documenti* che, data una collezione di documenti, restituisce una lista di documenti rilevanti, solitamente ordinati secondo un punteggio. Appartengono a questo gruppo i sistemi di individuazione e classificazione dell'argomento (topic) di un testo, e le applicazioni di ordinamento dei testi in base alla loro rilevanza;

Le tecnologie citate, anche se molto diverse tra loro, sono tutte suscettibili di un metodo di valutazione basato su uno standard di riferimento (*gold standard*). La prima fase per la creazione dello standard consiste nella definizione del task e di un formato di riferimento appropriato; quest'ultimo richiede a sua volta l'esplicitazione, tramite linee guida, di come e cosa annotare, lo sviluppo di strumenti in grado di supportare il progetto, e la validazione del processo di annotazione. Successivamente, insieme a questi strumenti, vengono realizzati e rilasciati dei corpora di addestramento annotati, che serviranno agli sviluppatori del sistema per implementare il task definito. In seguito, si procede alla validazione: il sistema, elaborando il corpus di prova, genera un insieme di risposte da confrontare con le risposte chiave dello standard di riferimento. La procedura di valutazione termina con il confronto tra il sistema candidato e lo standard di riferimento, insieme a test per la rilevanza statistica. Un esempio significativo di questo approccio si rivela nelle operazioni di *Named Entity*

Recognition[3]: in questo caso, lo standard di riferimento si configura come un testo integrato annotato con le entità da riconoscere.

Oltre al metodo che impiega un *gold standard*, esistono altri criteri di valutazione per i sistemi di analisi: uno di questi è la valutazione basata sulle caratteristiche del sistema. Tale orientamento caratterizza il metodo EAGLES, istituito dalla Comunità Europea e applicato agli strumenti per la traduzione (EAGLES 1996). La tecnica di valutazione EAGLES adotta un modello basato sui resoconti degli utenti: ad alcuni esperti del settore viene sottoposto il sistema di prova. Dopo averne esaminato le caratteristiche, gli esperti forniscono una serie di giudizi e compongono delle liste di controllo in cui sono elencate le caratteristiche critiche per ciascuna funzione o proprietà del componente da valutare.

Un approccio diverso consiste nella valutazione tramite componente incorporato, di solito adoperata quando un prodotto raggiunge lo stadio di prototipo. Questo metodo prevede che differenti versioni o implementazioni del componente da valutare vengano incorporate in un sistema più ampio, in modo da osservare le prestazioni nei vari casi. In questo scenario, sono fondamentali i concetti di *incorporabilità* (embeddability) e *portabilità* (portability): l'incorporabilità definisce il grado di facilità con cui è possibile inserire un componente in un sistema più ampio; la portabilità indica la possibilità di adattare un sistema per re-indirizzarlo verso nuovi compiti. La valutazione tramite incorporazione del componente si riscontra, per esempio, nei sistemi di riassunto automatico del testo: questi sono stati incorporati in un più largo sistema di *relevance assessment*[4] per il recupero di informazioni: la valutazione ha stimato quindi l'effetto che diverse tecniche di riassunto possono avere sulla velocità e l'accuratezza delle prestazioni (Mani et al. 1998). L'incorporazione può avvenire anche in esperimenti più complessi, dove è richiesta l'integrazione tra più componenti: è il

[3] Named Entity Recognition è un'attività di estrazione delle informazioni che individua gli elementi atomici nel testo e li classifica in categorie predefinite (come nomi di persone, organizzazioni, luoghi, quantità, percentuali, ecc.).

[4] Nella relevance assessment (o valutazione di rilevanza) vengono somministrati un testo e un argomento ad un sistema; quest'ultimo deve determinare la rilevanza del testo rispetto all'argomento.

caso del *question answering*[5], in cui il sistema deve rispondere alle domande estraendo le risposte da una vasta collezione di documenti. La valutazione sul *question answering* ha assunto modalità anche più complesse, per esempio sottoponendo il sistema a un test di *reading comprehension*: partendo da un racconto e da un serie di domande, il sistema doveva analizzare le domande, cercare nel testo le informazioni rilevanti e sintetizzare i risultati in una risposta.

Tutte le tipologie di valutazione presentate (il ricorso a uno standard di riferimento, la valutazione basata su caratteristiche e il metodo del componente incorporato) presentano vantaggi e svantaggi.

Nel caso del confronto con uno standard di riferimento, i vantaggi consistono nella riproducibilità della valutazione, nella capacità di generare risorse linguistiche e nella possibilità di supportare esperimenti di *machine learning*. Non mancano però alcuni aspetti negativi: anzitutto, l'impiego di un *gold standard* può generare risultati validi solo per lo specifico corpus coinvolto nella valutazione; infine, non bisogna sottovalutare gli ingenti costi richiesti dallo sviluppo delle infrastrutture necessarie alla valutazione (es. il gold standard).

La valutazione basata sulle caratteristiche è la più semplice ed economica da realizzare; diversamente da quanto avviene nello standard di riferimento, in cui la tecnologia è incentrata sui dati, le caratteristiche da valutare sono basate sul compito o sulla funzionalità del sistema. Questa peculiarità rende il metodo indipendente dal corpus, permettendo la valutazione di tipi di testo diversi. Ne deriva lo svantaggio che le risorse create sono troppo generiche per essere utilizzate da alcuni modelli di *machine learning*.

Infine, la valutazione tramite componente incorporato incentiva la diffusione della tecnologia *plug-and-play*, in modo da agevolare il processo di incorporazione. L'aspetto sconveniente di questo procedimento è la possibile influenza che il sistema generale può avere sul componente specifico.

[5] Il question answering consiste nel rispondere automaticamente a una domanda formulata in linguaggio naturale.

2.2 Valutazione dei sistemi di output

I sistemi di output restituiscono del linguaggio in uscita: fanno parte di questa categoria le tecnologie di riassunto, traduzione e generazione del linguaggio naturale. La valutazione può avvenire mediante misure intrinseche (valutando cioè il sistema in sé) e misure estrinseche (stimando l'efficienza e l'accettabilità del sistema).

Le **misure intrinseche** possono assumere varie forme: si può considerare l'output per sé stesso, oppure lo si può confrontare all'input o agli altri output. In ogni caso, la valutazione si concentra sul grado di qualità e di informatività dell'output.

La qualità è indipendente dall'informatività e determina quanto un testo sia coerente, ben formato e comprensibile a un parlante nativo. Varie misure possono essere adottate per giudicare la qualità: una delle più diffuse è la classificazione soggettiva, che individua le imprecisioni e le disfluenze nel testo in uscita, come gli errori grammaticali, la presenza di parole non tradotte, i difetti nello stile, la traduzione inadeguata dei nomi propri, la distruzione di ambienti strutturati quali liste o tabelle.

L'informatività stabilisce in che misura il testo di output preservi il contenuto informativo dell'input. La valutazione può configurarsi come un confronto tra il testo in uscita e il testo sorgente, con lo scopo di osservare la fedeltà nella resa dei risultati. L'impiego di questo metodo nello studio dei sistemi di *machine translation* ha evidenziato come a un testo sorgente possano corrispondere più traduzioni corrette, sebbene ciascuna differisca in qualche aspetto dal testo originale. Le discrepanze si accentuano quando l'informazione implicita nella lingua in ingresso deve essere resa esplicita nella lingua in uscita. Per esempio, nelle traduzioni dal cinese all'inglese molte informazioni possono perdersi, come la presenza di pronomi o argomenti del verbo. Le possibili soluzioni al problema sono due: la prima è di marcare il significato di ogni frase nel testo di input, elaborando una classificazione soggettiva della misura in cui il testo in uscita copre il testo in entrata. Ma è una metodologia che richiede sforzi costosi agli annotatori umani e che prevede un lavoro di classificazione soggettiva molto delicato. La seconda tecnica a cui ricorrere, più pratica ed efficiente, consiste nell'assegnare automaticamente un punteggio per valutare il valore informativo

del testo in uscita. L'informatività, intesa come misura dell'accuratezza nella copertura delle informazioni rispetto al testo sorgente, può essere misurata anche confrontando il risultato con altri output, prodotti da un computer o da un essere umano. Tuttavia, l'output di riferimento è potenzialmente incompleto, sia perché di uno stesso testo in ingresso possono essere forniti riassunti o traduzioni differenti, sia perché è difficile che il risultato del calcolatore coincida con l'elaborazione da parte di un essere umano. Di solito vi è comunque corrispondenza sulle parti più importanti da includere nel risultato; nei punti che suscitano disaccordo, si mettono in comune più output di riferimento. Alcuni progetti sperimentali più complessi creano delle raccolte miste di output di riferimento, dove si annoverano sia output realizzati da esseri umani sia output generati da calcolatori. Rispetto alle raccolte non miste, in queste valutazioni gli output umani risultano migliori in termini di precisione e chiarezza, al contrario degli output dei calcolatori, decisamente meno comprensibili.

In altri casi, l'output di riferimento si configura come un corpus di prova che contiene un insieme di esempi di output. È quanto accade nel lavoro di R. H. Robin (Robin 1994), dove sono introdotte due nuove misure: la robustezza, definita come la percentuale di frasi di output di prova che possono essere coperte senza aggiungere nuove conoscenze linguistiche al sistema, e la scalabilità, ossia la percentuale di nuovi concetti da aggiungere al sistema per coprire le frasi di prova. Sebbene l'uso di corpora di prova sia molto produttivo per la valutazione del sistema, la loro costruzione è complessa poiché i domini linguistici possono essere molto ampi. Un metodo più stimolante prevede degli output di riferimento per particolari categorie semantiche, come i nomi di persona, le informazioni sul tempo, i nomi di luoghi e altre terminologie specifiche. In questo quadro, il numero dei possibili output corretti si riduce notevolmente: ne risultano traduzioni più comprensibili e lineari.

Le **misure estrinseche** valutano il livello di efficienza con cui il sistema di output implementa le funzionalità di interesse per gli sviluppatori, i finanziatori o gli utenti finali. Esistono varie tipologie di misure estrinseche: un esempio sono le misure di *post-edit*, che contano la quantità di correzioni necessarie per una resa accurata e comprensibile dell'output; in pratica, individuano le aggiunte di parola, le cancellazioni, le trasposizioni. La resa accettabile dell'output può essere

determinata in base al testo di origine, a un output di riferimento o a un modello implicito che spiega in che termini un output può considerarsi adeguato. Un altro metodo di valutazione estrinseco si basa sull'efficienza nell'*esecuzione di istruzioni* da parte di un soggetto che legge un input; si tratta di un approccio adottato soprattutto quando i dati in uscita sono testi o manuali di natura didattica. Si prestano alla valutazione mediante misure estrinseche anche i test di *reading comprehension*, in cui un soggetto legge un output e risponde a delle domande, e i task di *relevance assessment*.

Quando un prodotto accede a uno stadio avanzato di sviluppo, è necessaria la valutazione dell'accettabilità da parte degli utenti finali; spesso questa si configura come un test che si accompagna a rilevamenti sulla velocità e le prestazioni del sistema di output. In tale scenario, le misure basate sulle caratteristiche si rivelano strumenti validi a esaminare vari tipi di applicazione, come quelle che supportano l'inserimento di input in molteplici formati, la pre e post revisione di un testo, l'aggiornamento del lessico, l'estensione della copertura linguistica del sistema, la gestione di diverse coppie di lingue, l'estensibilità a nuovi tipi di testo. Valutazioni di questo tipo richiedono considerevoli costi di sviluppo, controllo e manutenzione del sistema, oltre che spese per l'aggiornamento dei dizionari, la post-revisione del testo e la stampa.

2.3 Valutazione dei sistemi interattivi

I sistemi interattivi basano l'esecuzione del loro compito sull'interazione tra l'utente e il sistema. Alcuni esempi sono le applicazioni per la traduzione di documenti online, la ricerca basata sul Web, l'accesso telefonico a vari tipi di informazione. La valutazione dei sistemi interattivi da un lato richiede che l'utente e il sistema siano osservati e giudicati come una squadra, dall'altro non può prescindere dalla variabilità soggettiva delle entità coinvolte, e deve quindi sottoporre al test un numero di soggetti tale da offrire risultati statisticamente validi. Le misure tipiche sono il tempo e il costo necessari al compimento del lavoro, il grado di soddisfazione dell'utente e la qualità della soluzione proposta. Su questo tema si sono svolte varie tipologie di valutazione, a partire dagli anni

ottanta del Novecento, quando furono avviate le prime ricerche sulle interfacce dei database per il linguaggio naturale: l'obiettivo era di comparare l'uso del linguaggio naturale alle interfacce basate su menu (Whittaker e Walker 1989) e sul linguaggio SQL (Jarke et al. 1985). I sistemi si confrontavano sulla correttezza nella formulazione delle query e delle risposte, sul tempo necessario al compimento della richiesta e sulla capacità dell'utente di sfruttare alcune figure del linguaggio naturale, come le anafore, le ellissi e le congiunzioni.

La valutazione basata sul componente è stata adottata per l'analisi del sistema ATIS (Air Travel Information System), in cui lo standard di riferimento era composto da dati di prova pre-registrati e non prevedeva il coinvolgimento di utenti vivi; dunque la valutazione non forniva misure per prestazioni *end-to-end*.

Negli anni novanta nasce PARADISE (Walker et al. 1998), una rappresentazione per le interfacce di dialogo parlato che prende in esame più sistemi e definisce una funzione di valutazione della *performance*: alla sua definizione concorrono la misura della soddisfazione degli utenti, la difficoltà del compito da portare a termine e il costo del dialogo.

Un approccio innovativo proviene dalla valutazione di interfacce telefoniche di conversazione per le informazioni meteorologiche (Polifroni et al. 1998). I dati vengono registrati, trascritti e valutati con frequenza quotidiana; è una scelta che facilita notevolmente la gestione di grandi quantità di dati e ottimizza i tempi del lavoro. Le misure caratteristiche in questo tipo di valutazione sono la correttezza semantica e di generazione, e il numero di errori di parole.

Un'altra metodologia, affermatasi prevalentemente in Europa, prende le mosse dall'uso estensivo degli esperimenti *Wizard of Oz*[6] nella progettazione dei sistemi interattivi (Bernsen e Dybkjaer 1998). Durante la fase di progettazione del sistema, gli sviluppatori possono valutare come gli utenti rispondono a determinate sollecitazioni e quali termini o costruzioni sintattiche prediligono nelle diverse situazioni.

[6] In un esperimento Wizard of Oz dei soggetti interagiscono con un sistema informatico che si finge autonomo, ma che in realtà viene gestito, totalmente o in parte, da un essere umano nascosto.

Lo sviluppo delle applicazioni interattive richiede una valutazione più complessa rispetto alle altre tipologie di sistema; vi è un bisogno urgente di ricerca in quest'area sia per rendere gli esperimenti più semplici, economici e riutilizzabili, sia per fornire adeguate misure diagnostiche utili alla maturazione dei sistemi.

A conclusione dello sguardo sui diversi metodi valutativi, è evidente come la valutazione sia un'attività sociale, in quanto: crea una comunità che confronta varie tecnologie mediante criteri condivisi; promuove una sana competizione che incoraggia al miglioramento dei sistemi; infine, produce infrastrutture e risorse riutilizzabili. Tra le tecniche di valutazione descritte, quelle relative ai sistemi di analisi hanno subìto uno sviluppo molto più rapido. Lo studio dei sistemi interattivi e di output è invece più problematico, sia perché richiede ricerche approfondite sui costi di interazione con l'utente, sia perché deve considerare le questioni relative a portabilità e incorporabilità. In generale, i criteri di valutazione devono continuamente adeguarsi all'accelerazione insistente del ciclo di vita della tecnologia, che ad oggi esige flussi di feedback sempre più frequenti.

3. Il progetto Evalita

Evalita è un'iniziativa che promuove la valutazione di strumenti per l'elaborazione automatica del linguaggio. Lo scopo di questo progetto, nato nel 2007, è di favorire lo sviluppo di risorse e strumenti utili al Natural Language Processing e alle scienze del linguaggio, concentrando l'attenzione sul trattamento della lingua italiana. La valutazione degli strumenti coinvolti in Evalita avviene in uno scenario condiviso, in cui diversi sistemi possono essere esaminati e comparati in maniera coerente con misure e procedure per la valutazione formale. A ciascun partecipante viene fornito un *training corpus* (ovvero un corpus di addestramento) e un corpus di prova, a sua volta suddiviso in *dev set*, da usare durante lo sviluppo del sistema, e *test set*, contenente dati di prova per la valutazione. Inoltre, i partecipanti dispongono delle metriche di valutazione condivise con cui valutare i propri sistemi. Il progetto Evalita ha ottenuto fin dall'inizio un buon riscontro, sia in termini di numero dei partecipanti che di qualità dei risultati, e ad oggi conta tre edizioni (avvenute rispettivamente nel 2007, 2009 e 2011). Un importante effetto collaterale delle campagne di valutazione organizzate da Evalita consiste nella possibilità di riutilizzare i dati di prova, che sono resi disponibili alla comunità scientifica.

3.1 Le tre edizioni di Evalita

Le tre edizioni di Evalita presentano dei tratti distintivi, che documentano i progressi raggiunti dal progetto. La prima edizione, svoltasi nel 2007, è nata con la convinzione che la diffusione di *task* e pratiche condivise di valutazione fosse un passo cruciale per lo sviluppo del trattamento automatico del linguaggio. Fin dal primo momento, l'iniziativa ha ricevuto il sostegno e la partecipazione di molte istituzioni, sia del mondo accademico che di quello industriale. La campagna di valutazione del 2007 ha però coinvolto soltanto risorse di tipo testuale, tralasciando le tecnologie di elaborazione del parlato. Sono stati analizzati cinque *tasks*: Part of Speech Tagging, Parsing, Word Disambiguation Sense, Temporal Expression Recognition and Normalization e Named Entity Recognition.

L'edizione successiva, del 2009, ha introdotto la valutazione delle tecnologie per il parlato, affiancando ai *text tasks* una serie di *speech tasks*. Questi sono: Connected Digits Recognition, Spoken Dialogue Systems Evaluation e Speaker Identity Verification. Lo stimolo a integrare il progetto con aspetti della *speech technology* proviene dall'interesse per il modello del NIST (National Institute of Standards and Technology), impiegato nelle campagne di valutazione statunitensi. È un modello molto valido, in quanto comprende la definizione di standard e misure di valutazione, la collezione di database, l'organizzazione di campagne di valutazione aperte e la definizione di nuovi tasks e obiettivi. Inoltre, il modello NIST ha creato un ambiente in cui ospitare una sana competizione tra le varie imprese, al fine di migliorare le potenzialità dei sistemi. Un progetto simile non è mai giunto in Europa, sia a causa della minore attenzione verso i sistemi di parlato, sia per lo stato frammentario della comunità di ricercatori in questo campo. Ispirata dal modello NIST, la seconda edizione di Evalita ha voluto colmare le lacune dell'edizione precedente nell'ambito delle tecnologie del parlato, pur considerando i rischi di un progetto così innovativo: infatti, lo stato poco avanzato della ricerca in Italia ha comportato qualche difficoltà nel reperimento delle risorse in lingua italiana.

La terza ed ultima edizione, organizzata nel 2011, ha riproposto la partizione in *text tasks* e *speech tasks*, confermando il successo delle edizioni precedenti. I *task* testuali valutati sono: Parsing, Domain Adaptation, Named Entity Recognition on Transcribed Broadcast News, Cross-document Coreference Resolution, Anaphora Resolution, Super Sense Tagging, Frame Labeling over Italian Texts e Lemmatisation. I *task* per il discorso parlato comprendono: Automatic Speech Recognition, Forced Alignment on Spontaneous Speech e Voice Applications on Mobile. La presente relazione non prenderà in esame tutti i tasks citati, ma si concentrerà sui compiti di annotazione semantica in Evalita 2011, ovvero Frame Labeling over Italian Texts e Super Sense Tagging.

3.2 Valutazione di task semantici in Evalita 2011

I task analizzati nella relazione consistono nell'annotazione semantica di risorse testuali. Un'annotazione di questo tipo mira a rendere la semantica delle

informazioni esplicita ed accessibile al computer. L'annotazione semantica permette di superare l'ambiguità del linguaggio naturale, esprimendo i concetti e la loro rappresentazione computazionale in un linguaggio formale. Inoltre, addestrando un computer su come i dati sono correlati e su come queste relazioni possono essere elaborate automaticamente, diviene possibile implementare operazioni di ricerca complesse. Rispetto al *Part-of-speech* tagging (che associa ad ogni parola la sua categoria lessicale) e al tagging sintattico (che etichetta le unità sintattiche del testo), il tagging semantico agisce a un livello più profondo: arricchisce i dati con informazioni legate al loro significato.

La campagna Evalita 2011 ha documentato la valutazione dei sistemi per tasks semantici, in particolare basati sui compiti di Frame Labeling over Italian Texts e Super Sense Tagging. Nel primo caso, i sistemi hanno marcato i ruoli semantici instanziati nel testo, ovvero le funzioni semantiche svolte dai sintagmi nell'evento espresso dal verbo. Il secondo tipo di task ha invece previsto la classificazione delle parole in base a una tassonomia di categorie concettuali definita da WordNet.

In entrambi i casi, l'annotazione a questo livello costituisce uno strumento prezioso per l'estrazione di conoscenza semantica dal testo; ciò si rivela fondamentale nella costruzione formale di ontologie e nei compiti più complessi di Semantic Search, Information Extraction e Information Retrieval.

3.2.1 Super Sense Tagging

Il Super Sense Tagging (SST) è un task semantico che prevede l'annotazione di ogni entità lessicale nel testo (nomi, verbi, aggettivi e avverbi), facendo riferimento a un ordinamento semantico definito dalle classi lessicografiche di WordNet. WordNet (Fellbaum 1998) è un ampio database lessicale in lingua inglese che raggruppa sostantivi, verbi, aggettivi e avverbi in insiemi di concetti lessicali (*synset*). Ogni synset esprime un concetto distinto ed è collegato agli altri tramite una rete di relazioni semantiche e lessicali. I synset sono organizzati in domini diversi, in base alla categoria sintattica e alla coerenza semantica: ovvero, a ogni parola contenuta in un synset viene assegnata una delle 45 classi

lessicografiche di WordNet. Tali classi sono chiamate SuperSenses poiché raggruppano grandi insiemi di synset, e si suddividono in 26 per i nomi, 15 per i verbi, 3 per gli aggettivi e una per gli avverbi. Di seguito è riportato l'elenco dei SuperSenses:

Id	**SuperSense**	**Description**
00	adj.all	all adjective clusters, used for all simple adjectives
01	adj.pert	Relational adjectives (pertainyms), adjectives that are related with nouns
02	adv.all	all adverb
03	noun.Tops	unique beginner for nouns, nouns that appear at top level
04	noun.act	nouns denoting acts or actions
05	noun.animal	nouns denoting animals
06	noun.artifact	nouns denoting man-made objects
07	noun.attribute	nouns denoting attributes of people and objects
08	noun.body	nouns denoting body parts
09	noun.cognition	nouns denoting cognitive processes and contents
10	noun.communication	nouns denoting communicative processes and contents
11	noun.event	nouns denoting natural events
12	noun.feeling	nouns denoting feelings and emotions
13	noun.food	nouns denoting foods and drinks
14	noun.group	nouns denoting groupings of people or objects
15	noun.location	nouns denoting spatial position
16	noun.motive	nouns denoting goals

17	noun.object	nouns denoting natural objects (not man-made)
18	noun.person	nouns denoting people
19	noun.phenomenon	nouns denoting natural phenomena
20	noun.plant	nouns denoting plants
21	noun.possession	nouns denoting possession and transfer of possession
22	noun.process	nouns denoting natural processes
23	noun.quantity	nouns denoting quantities and units of measure
24	noun.relation	nouns denoting relations between people or things or ideas
25	noun.shape	nouns denoting two and three dimensional shapes
26	noun.state	nouns denoting stable states of affairs
27	noun.substance	nouns denoting substances
28	noun.time	nouns denoting time and temporal relations
29	verb.body	verbs of grooming, dressing and bodily care
30	verb.change	verbs of size, temperature change, intensifying, etc.
31	verb.cognition	verbs of thinking, judging, analyzing, doubting
32	verb.communication	verbs of telling, asking, ordering, singing
33	verb.competition	verbs of fighting, athletic activities
34	verb.consumption	verbs of eating and drinking
35	verb.contact	verbs of touching, hitting, tying,

		digging
36	verb.creation	verbs of sewing, baking, painting, performing
37	verb.emotion	verbs of feeling
38	verb.motion	verbs of walking, flying, swimming
39	verb.perception	verbs of seeing, hearing, feeling
40	verb.possession	verbs of buying, selling, owning
41	verb.social	verbs of political and social activities and events
42	verb.stative	verbs of being, having, spatial relations
43	verb.weather	verbs of raining, snowing, thawing, thundering
44	adj.ppl	participial adjectives

L'obiettivo del tagging è di assegnare un SuperSense appropriato per ogni token o espressione polirematica. Nella classificazione non sono compresi i verbi modali e di supporto, in quanto non sono portatori di informazione semantica. L'uso di questo task si rivela molto proficuo nelle applicazioni che richiedono la comprensione e la conoscenza del mondo, come il recupero e l'estrazione di informazioni semantiche (*Information retrieval* e *Information extraction*) o il *Question answering*. Il SuperSense Tagging si colloca a metà tra i task di Named Entity Recognition e Word Sense Disambiguation: è un'estensione del primo, perché usa un insieme più ampio di categorie semantiche, ed è più semplice e pratico rispetto al secondo, in quanto coinvolge classi di senso meno specifiche.

La valutazione del task di SST in Evalita 2011 ha previsto l'impiego delle 45 categorie di SuperSense. Il compito è stato suddiviso in due sottotask: un *closed subtask*, in cui valutare i sistemi usando solo il training corpus fornito da Evalita, e un *open subtask*, dove è stato consentito il ricorso a risorse esterne differenti dal training corpus. Il dataset proviene da una revisione del ISST-SST corpus, una risorsa di 300.000 tokens realizzata nel 2010 per scopi di ricerca (Montemagni et al. 2003). La raccolta è stata aggiornata appositamente per gli scopi di Evalita

2011, tramite il completamento delle annotazioni mancanti per 24.000 token, e lo sviluppo di nuove strategie di tagging per le espressioni polirematiche. Da questa versione aggiornata sono stati prelevati 256.000 token impiegati come dati per l'addestramento e lo sviluppo dei sistemi. Il corpus di prova, invece, è stato creato a partire da una porzione dell'ISST non aggiornata cui si è aggiunta una collezione di frasi (circa 20.000 tokens) estratte dalla versione italiana di Wikipedia, poi annotate e revisionate a mano. Tutti i dati sono stati codificati in UTF-8 e organizzati in documenti; ciascun documento contiene una serie di frasi, in cui ogni riga è occupata da un token. Affianco a ciascun token sono stati inseriti quattro campi, prodotti con strumenti automatici ma revisionati manualmente: FORM, LEMMA, PoS e SuperSense. Di seguito, un esempio di frase annotata:

FORM	**LEMMA**	**PoS**	**SuperSense**
Un	un	RIms	O
Incendio	incendio	Sms	B-noun.event
,	,	FF	O
che	che	PRnn	O
si	si	PC3nn	O
sarebbe	essere	VAd3s	O
sviluppato	sviluppare	Vpsms	B-verb.creation
per	per	E	O
cause	causa	Sfp	B-noun.motive
accidentali	accidentale	Anp	B-adj.all
,	,	FF	O
ha	avere	VAip3s	O
gravemente	gravemente	B	B-adv.all
danneggiato	danneggiare	Vpsms	B-verb.change
a	a	E	O
Fiano	fiano	SP	B-noun.location
,	,	FF	O
uno	uno	RIms	O
chalet	chalet	Smn	B-noun.artifact
di	di	E	O
proprietà	proprietà	Sfn	B-noun.possession

di	di	E	O
Umberto	umberto	SP	B-noun.person
Agnelli	agnelli	SP	I-noun.person

Le misure di valutazione impiegate sono *tagging accuracy*, cioè la percentuale di tokens correttamente classificati rispetto al numero totale di tokens, e *F1-measure*, ovvero la media armonica pesata tra precisione e recall:

$$\text{Tagging accuracy} = \frac{\text{correctly detected tokens}}{\text{total number of tokens}} \cdot 100;$$

$$\text{F1-measure} = \frac{\text{precision} \cdot \text{recall}}{\text{precision+recall}} \cdot 2;$$

Hanno partecipato alla valutazione del task di SuperSense Tagging l'Università di Pisa e l'Università di Bari, sottoponendo i loro sistemi a quattro turni di prova. Si è verificata una distinzione nell'esecuzione dei sottotask: mentre la squadra di Pisa ha preso parte solo al closed subtask, l'Università di Bari ha aderito a entrambi. Il sistema dell'Università di Pisa ha elaborato i dati tramite un classificatore basato sull'algoritmo di Maximum Entropy, per frammentare il testo in chunks, e un algoritmo di programmazione dinamica, per selezionare le sequenze di tag con probabilità più alta. Il sistema ha estratto tre tipi di caratteristiche dai dati: le caratteristiche degli attributi, relative agli attributi dei tokens circostanti; le caratteristiche locali, riguardanti la forma della parola e il suo contesto; le caratteristiche globali, ovvero le proprietà a livello del documento.

L'Università di Bari ha partecipato alla campagna con due sistemi, fondando la classificazione su Support Vector Machine. Nell'open subtask è stato necessario impiegare alcune caratteristiche di un WordSpace semantico, per arginare il problema della sparsità dei dati[7]. Il modello del WordSpace colloca le parole e i concetti in uno spazio matematico, in cui i concetti con significati simili sono vicini tra loro. Le regole di utilizzo di una parola in un determinato contesto definiscono il suo significato e, pertanto, le parole che condividono un contesto si considerano semanticamente simili.

[7] La sparsità dei dati indica la mancanza di dati annotati affidabili coi quali costruire i modelli dei fenomeni linguistici.

I punteggi migliori raggiunti dai sistemi hanno valori molto vicini tra loro, con una leggera discrepanza nell'esecuzione dei sottotasks: mentre l'Università di Pisa ha ottenuto risultati migliori nel closed subtask, la superiorità nell'open subtask spetta all'Università di Bari. È importante anche sottolineare la diversa influenza che due porzioni di testo da cui è stato ricavato il corpus di prova hanno esercitato sull'elaborazione: il modello addestrato sul traning corpus ISST-SST ha mostrato una notevole abilità nel far fronte a un dominio diverso senza bisogno di strategie di adattamento. Questo comportamento è stato riscontrato nei sistemi dell'Università di Pisa, ma non nei sistemi dell'UniBA, dove le prestazioni hanno subìto un leggero declino.

3.2.2 Frame Labeling over Italian Texts

Nel task di Frame Labeling over Italian Texts (FLaIT) i sistemi hanno dovuto individuare, in una frase in lingua italiana, il frame semantico evocato da un predicato (*target*) e i principali ruoli semantici (*frame elements*) esplicitamente menzionati. In particolare, il compito consisteva nel riconoscere le parole e le frasi che evocano frames semantici del tipo definito nel progetto FrameNet (Baker et al. 1998), e i loro dipendenti semantici, che di solito (ma non sempre) coincidono con i dipendenti sintattici: questo task viene definito Semantic Role Labeling (SRL). L'obiettivo della valutazione era di presentare modelli di rappresentazione, algoritmi e metodi di inferenza induttiva che supportassero il SRL. Si tratta di un task nuovo in questo scenario, poiché viene sottoposto per la prima volta alle metriche di Evalita. Le esperienze di valutazione precedenti hanno impiegato due corpora in lingua inglese per l'addestramento dei sistemi: PropBank (Palmer et al. 2005) e FrameNet. L'interesse di Evalita per il Frame Labeling mira allo sviluppo di risorse simili a FrameNet per l'italiano, che al momento sono in fase di crescita nel progetto iFrame[8].

FrameNet è un'iniziativa promossa dall'International Computer Science Institute di Berkeley, California, che si occupa dell'annotazione di risorse elettroniche mediante frames semantici, teoria derivata dal lavoro di Charles J. Fillmore e colleghi. Un *frame* fornisce il modello di un evento, uno stato o una situazione

[8] Sito del progetto: http://sag.art.uniroma2.it/iframe/doku.php.

evocato da un'unita lessicale denominata *target*. Alla realizzazione del frame concorrono una serie di ruoli semantici, o *frame elements*, i quali descrivono le relazioni semantiche tra il verbo e i suoi argomenti. Il compito di FrameNet è di definire l'insieme dei possibili frames e di annotare le frasi, in modo da istanziare sintatticamente i frame elements attorno alla parola che evoca il frame.

Prendendo come esempio il frame *Commerce_pay*, in cui un compratore paga per ricevere un prodotto, sono coinvolti i seguenti ruoli semantici:

1. un *Buyer*, ovvero il compratore;
2. *Goods*, qualsiasi prodotto, anche immateriale, venga ceduto al compratore;
3. *Money*, il compenso dato dal compratore per ottenere il prodotto;
4. *Seller*, il venditore che fornisce il prodotto;
5. e infine *Rate*, che indica la possibilità di frazionare il pagamento in più unità.

Tutti questi frame elements appartengono all'insieme degli elementi nucleari del frame, e rappresentano pertanto i ruoli semantici fondamentali del modello evocato. Esistono anche degli elementi non nucleari che possono essere eventualmente aggiunti all'annotazione: spesso servono a descrivere fattori circostanziali come il tempo, il luogo, lo scopo o la maniera in cui si svolge la situazione, e si presentano sotto forma di sintagmi preposizionali o avverbi. Da un punto di vista grammaticale, questi elementi non possono essere né soggetto né oggetto del verbo target. Il database di FrameNet include più di mille frames, legati fra loro tramite una rete di relazioni semantiche: queste possono stabilire connessioni tra le strutture semantiche più generali e quelle più specifiche, oltre che fornire strumenti per la riflessione sugli eventi e sulle azioni. L'obiettivo dell'iniziativa è di creare e rendere disponibile un repertorio di situazioni semantiche che possano essere impiegate nei sistemi di elaborazione del linguaggio naturale. Il ricorso ai frames si rivela estremamente utile nei task di comprensione testuale o di disambiguazione, in cui è necessario associare in modo automatico una o più strutture semantiche alle frasi del testo.

Per procedere alla valutazione, il task FLaIT è stato suddiviso in tre *subtasks*: Frame Prediction (FP), Boundary Detection (BD) e Argument Classification (AC); in FP, data un'unità lessicale ambigua, il sistema deve riconoscere e

assegnare il frame corretto. Il formato di annotazione presenta un token per riga ed ogni riga si compone di quattro campi: *token id*, ovvero l'identificatore univoco del token nell'ambito della frase; la *forma* flessa del token; *Pos tag*, che indica la categoria sintattica del token; e infine un ultimo campo dedicato all'assegnazione del target e dei ruoli semantici. Di seguito, un esempio di frase in cui è stato individuato il target (*pagare*) per il frame *Commerce_pay*.

1	Ciò	PD	-	
2	premesso	A		
3	,	FF	-	
4	intende	V	-	
5	la	RD	-	
6	Commissione	SP		
7	modificare	V		
8	questo	DD	-	
9	stato	S	-	
10	di	E	-	
11	cose	S	-	
12	,	FF	-	
13	di	E	-	
14	modo	S	-	
15	che	CS	-	
16	anche	B	-	
17	i	RD	-	
18	funzionari	S		
19	UE	SP	-	
20	debbano	VM		
21	pagare	V	Commerce-pay	Target
22	l'	RD	-	
23	imposta	S		
24	sui	EA	-	

25	redditi	S	-	
26	nel	EA	-	
27	paese	S	-	
28	in	E	-	
29	cui	PR	-	
30	lavorano	V		
31	?	FS	-	

In BD e AC bisogna rispettivamente individuare tutti gli argomenti semantici di un frame e assegnare loro un frame element, data l'unità lessicale. I confini dei ruoli semantici si individuano mediante l'assegnazione di un frame element alle unità lessicali che lo realizzano. Di seguito, è riproposta la stessa frase, di cui ora sono definiti anche i confini dei ruoli e i rispettivi frame elements:

1	Ciò	PD	-	-	
2	premesso	A	-	-	
3	,	FF	-	-	
4	intende	V	-	-	
5	la	RD	-	-	
6	Commissione	SP	-	-	
7	modificare	V	-	-	
8	questo	DD	-	-	
9	stato	S	-	-	
10	di	E	-	-	
11	cose	S	-	-	
12	,	FF	-	-	
13	di	E	-	-	
14	modo	S	-	-	
15	che	CS	-	-	
16	anche	B	-	-	
17	i	RD	-	Buyer	

18	funzionari	S	-	Buyer
19	UE	SP	-	Buyer
20	debbano	VM	-	-
21	pagare	V	Commerce-pay	Target
22	l'	RD	-	Money
23	imposta	S	-	Money
24	sui	EA	-	Money
25	redditi	S	-	Money
26	nel	EA	-	Place
27	paese	S	-	Place
28	in	E	-	Place
29	cui	PR	-	Place
30	lavorano	V	-	Place
31	?	FS	-	-

L'insieme di dati forniti per l'addestramento del sistema nasce dalla fusione di due diverse risorse annotate: la prima è un set sviluppato dalla Fondazione Bruno Kessler (Tonelli et al. 2008) , formato da 605 frasi e 1074 ruoli annotati a livello sintattico e semantico; la seconda risorsa è un set di 650 frasi e 1763 ruoli, creato dall'Istituto di Linguistica Computazionale di Pisa (Lenci et al.): si tratta dell'ISST-TANL Corpus, risultato della revisione di un sottoinsieme della ISST (Montemagni et al. 2003), e dell'integrazione con frames semantici. Il set di addestramento risultante comprende 1255 frasi per 38 frames. Il test set su cui effettuare gli esperimenti è stato ottenuto mediante l'allineamento automatico tra il lessico inglese di FrameNet e l'italiano: per fare ciò, l'Università di Roma, Tor Vergata, ha impiegato Moses (Basili et al. 2009), uno strumento open-source di *statistical machine translation* (SMT), ovvero traduzione automatica su base statistica.

L'allineamento lessicale tra due lingue consiste nel trasferimento di informazione semantica da una lingua all'altra. L'idea che i corpora annotati per la lingua inglese possano essere utilizzati per trasferire informazione semantica su un'altra

lingua induce allo sviluppo di risorse annotate mediante frames per lingue europee diverse dall'inglese. Se si considera un corpus bilingue, avente l'inglese come lingua sorgente (*E*) e l'italiano come lingua target (*T*), l'allineamento richiede dapprima la selezione delle coppie di frasi (s_E, s_T) in grado di realizzare uno specifico frame *f*, e poi l'annotazione della frase s_T in base a *f*. Il procedimento si basa sull'impiego di uno strumento di SMT, per individuare le traduzioni candidate, e su un modello di selezione delle frasi. Una volta determinate le coppie di frasi idonee, i dati vengono rielaborati per definire i confini dei ruoli semantici. Una coppia di frasi allineate è idonea quando costituisce un esempio valido per un frame *f*, ovvero quando entrambe le frasi esprimono la specifica informazione semantica veicolata da *f*.

L'insieme di esperimenti di FLaIT eseguiti sui sistemi ha seguito un approccio strutturato, per cui i dati di prova sono stati presentati in maniera incrementale, con un numero di dettagli crescente a ogni turno. Nel primo turno le frasi sono state contrassegnate solo dalla parola target, senza alcuna informazione sul frame evocato; nel secondo turno si è aggiunto il compito di individuazione del frame corretto, ma si è tralasciata la definizione dei confini; il terzo turno ha infine previsto anche l'annotazione degli argomenti e la presentazione di informazioni esatte sui confini. Sebbene il livello di dettaglio dei dati sia stato reso visibile in modo progressivo, in ciascuna fase è stata richiesta l'esecuzione di tutti i sottotask menzionati, in modo da valutare quanto gli errori di etichettatura iniziali potessero incidere sulla qualità delle successive fasi di annotazione, e quanto i vari sistemi fossero agevolati dall'avere a disposizone i vari livelli di informazione (frame e confini dei frame elements).

Le misure di valutazione adottate sono precisione e recall, e si basano sul numero di *true positives* riscontrati nei sottotask. Nel compito di Frame Prediction, è considerata *true positive* ogni coppia frase-predicato per cui è fornito il frame corretto. In Boundary Detection, i *true positives* sono gli argomenti semantici di cui vengono determinati con precisione i confini. In Argument Classification, si ritengono *true positives* gli argomenti a cui è stato assegnato il ruolo semantico corretto, e *false positives* gli argomenti la cui annotazione non corrisponde all'etichetta nel corpus di riferimento. Gli argomenti non annotati sono *false negatives*. La precisione in quest'ultimo sottotask è data dal numero di true

positives diviso per la somma di true positives e false positives; il recall si ottiene dividendo la quantità di true positives per la somma di true positives e false negatives.

La campagna di valutazione per il task di FLaIT ha ricevuto la partecipazione di quattro sistemi, riconducibili a due diverse istituzioni: CELI e l'Università di Roma, Tor Vergata. I due FLaIT CELI Systems hanno sottoposto le frasi in input a un *parser* esistente e hanno basato la propria analisi su una combinazione di regole di dipendenza e tecniche di *machine learning*. L'implementazione di metodi per i sottotask di Frame Labeling e Boundary Detection è stata sviluppata dagli autori del CELI System proprio in occasione della partecipazione a Evalita 2011. I sistemi proposti dall'Università di Roma sono due: lo Structured Learning SRL system, che si fonda sul concetto di *structured learning*[9], e il Semi-Supervised SRL system, che adotta un'architettura ibrida per i task di Boundary Detection e Argument Classification.

Nel sottotask di Frame Labeling i sistemi del CELI hanno ottenuto il 73.93% di precisione e il 65.09% di recall nel sottotask di Frame Labeling, mentre l'Università di Roma ha raggiunto, sia in precisione che in recall, il punteggio di 80.82% con il primo sistema, e 78.62% con il secondo. In Boundary Detection, sia nel primo che nel secondo turno, i sistemi CELI hanno ottenuto un grado di precisione attorno al 40% e una recall di circa il 20%; i sistemi di Roma hanno ottenuto una precisione di più del 60% e una recall intorno al 50%. Nel sottotask di Argument Classification, durante i primi due turni, i sistemi del CELI hanno mostrato un punteggio di precisione tra il 27% e il 36% e un punteggio di recall tra il 14% e il 19%, mentre i sistemi di Roma hanno ottenuto un grado di precisione e di recall tra il 33% e il 55%. Le prestazioni sono notevolmente migliorate nel terzo turno, dove i sistemi del CELI hanno raggiunto una precisione del 75% e una recall del 40%, e i sistemi di Roma hanno ottenuto il 70% per entrambe le misure.

I punteggi migliori ottenuti dai sistemi sono abbastanza alti e simili tra loro, in parte a causa del numero esiguo di frames da identificare: i partecipanti hanno

[9] Structured learning è un sottocampo del machine learning e riguarda programmi che imparano a mappare i dati in input su output arbitrariamente complessi.

avuto a disposizione 105 diverse unità lessicali da assegnare a 36 frames. Inoltre, il task da portare a termine era piuttosto semplice: i target sono stati ampiamente marcati nel corpus di prova, e quindi i sistemi dovevano solo scegliere il frame giusto tra quelli attestati nel training corpus per l'unità lessicale in questione. Valori molto simili sono stati raggiunti nella misura della precisione, mentre i punteggi di recall presentano notevoli differenze da un sistema all'altro. Analizzando i tre turni di prova cui sono stati sottoposti i sistemi, non si sono riscontrate differenze sostanziali nel Boundary Detection: ciò significa che conoscere il frame evocato dal target non aiuta a identificare i confini dei suoi frame elements. Nel sottotask di Argument Classification il punteggio migliora leggermente nella seconda esecuzione e ha un notevole incremento nella terza, a dimostrazione di come l'individuazione del tipo di frame e dei confini dei ruoli semantici sia utile per l'assegnazione dei frame elements.

Lo svolgimento della campagna ha messo in luce una varietà di metodi per lo sviluppo di tecnologie di Semantic Role Labeling, a partire dall'uso di regole di dipendenza fino a criteri probabilistici e discriminanti. In generale, tutti i sistemi hanno mostrato una capacità di precisione talmente valida da poter essere paragonata alle tecnologie per la lingua inglese, che vanta una gamma di risorse certamente più ricca rispetto all'italiano.

4. Preparazione del test set per FLaIT a Evalita 2011

La presente relazione nasce dall'attività di stage svolta tra settembre 2011 e gennaio 2012, concentrata sulla campagna di valutazione Evalita 2011. L'esperienza di tirocinio ha previsto la partecipazione a un progetto finalizzato all'annotazione di frasi in lingua italiana per mezzo di frames semantici in stile FrameNet.

Il metodo di FrameNet ha rivestito un ruolo centrale nel progetto, che può essere diviso in due fasi distinte. Nella prima fase, l'annotazione ha condotto alla realizzazione di un corpus annotato da utilizzare come test set nel task di Frame Labeling over Italian Texts in Evalita 2011. La seconda fase ha previsto l'utilizzo del software di annotazione SALTO su un campione di frasi di dominio giornalistico, al fine di arricchire la ISST (Italian Syntactic-Semantic Treebank) con nuove istanze di frames.

Il lavoro di annotazione si è svolto presso l'Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR a Pisa. Il primo stadio del progetto si è concentrato sull'addestramento, guidato dalla dott.ssa Venturi, riguardo allo schema di annotazione e al software da adoperare. Una volta acquisite le conoscenze necessarie, ho proceduto alla validazione e alla revisione manuale di un campione di frasi annotate automaticamente da Moses.

Nell'ambito del progetto descritto, Moses è stato utilizzato per l'annotazione di una porzione di testo allineata tra inglese (s_E) e italiano (s_I) del corpus Europarl. Moses permette la creazione di tabelle di *phrase translation* (PT), in cui sono riportati tutti i segmenti di testo s_E che hanno una traduzione in s_I. L'output dell'allineamento statistico consiste in una serie di coppie di segmenti (*es*, *is*), ponderate secondo una probabilità che descrive una corrispondenza molti a molti tra un elemento semantico inglese e alcuni segmenti *is* in s_I. Le coppie possono essere formate da due parole o da un segmento inglese affiancato a un segmento italiano più lungo.

Data una coppia di frasi (s_E, s_I), un elemento α (ad esempio, un ruolo semantico) e il segmento inglese che esprime il ruolo $s_E(\alpha)$, l'allineamento in Moses segue quattro fasi:

1. *Rank Phase*: vengono classificati tutti i segmenti tradotti collegati ad almeno una parola in $s_E(\alpha)$. L'ordinamento avviene secondo un determinato criterio, come ad esempio la lunghezza;
2. *Collect Phase*: si esaminano le coppie contenute nella tabella di PT. Si selezionano quindi i candidati per tutti i token in $s_E(\alpha)$, fino a coprire interamente il segmento inglese con una traduzione;
3. *Boundary Detection Phase*: si procede alla definizione dei confini dei ruoli semantici nei segmenti in italiano;
4. *Post-Processing Phase*: si perfezionano i confini dei ruoli mediante l'applicazione di euristiche basate sull'intera frase.

Moses raggiunge valori di accuratezza soddisfacenti nell'allineamento tra inglese e italiano: il *Perfect Matching*, ovvero il perfetto riconoscimento del frame e la corretta assegnazione dei ruoli semantici, ammonta al 66,88% di accuratezza.

La definizione del test set per FLaIT ha previsto l'impiego di Moses al fine di allineare il lessico inglese di FrameNet alla lingua italiana, individuando le frasi candidate per ciascuno dei 38 frames già definiti nel training dataset. Il set risultante comprende 318 frasi e 560 argomenti.

La prima parte del lavoro si è concentrata su un ciclo di annotazione semi-automatica di frasi di dominio giuridico. Dati 38 frames, è stato richiesto di rintracciare le istanze dei frames nelle frasi e di revisionare manualmente l'individuazione dei ruoli semantici. La procedura si è svolta a partire da un primo repertorio di cento frasi per frame, pre-annotate automaticamente da Moses, e considerate come potenziali istanze per un determinato frame. Successivamente, è stato richiesto di selezionare dalla collezione almeno dieci frasi appropriate per ciascuna situazione semantica. La selezione è stata condotta in base a due criteri fondamentali: la scelta è stata orientata solo su target di tipo verbale, tralasciando le altre parti del discorso, e sono state preferite frasi che nel target non presentassero lo stesso lemma. Per risolvere gli eventuali casi di incertezza, vi era a disposizione il medesimo training dataset fornito ai sistemi partecipanti alla competizione. Nell'eventualità in cui non si riuscisse a rintracciare il numero minimo di frasi richieste per ciascun frame, era possibile ottenere nuovi insiemi di frasi pre-annotate su cui effettuare la ricerca. Dopo aver selezionato le dieci frasi

idonee per ogni frame, è stata revisionata l'annotazione dei frame elements: le frasi sono state quindi riviste e validate manualmente, correggendo l'assegnazione dei ruoli semantici in caso di errore. Il risultato finale comprende 318 frasi ricondotte a 38 frames diversi, per un totale di 105 unità lessicali e 560 argomenti semantici identificati.

Nella seconda parte del progetto è stato impiegato il software di annotazione SALTO per individuare nuove istanze di frames in un corpus di dominio giornalistico. Lo scopo della ricerca era di estendere la ISST.

SALTO è uno strumento grafico sviluppato dal progetto SALSA che supporta l'annotazione manuale di corpora di testo. È stato progettato appositamente per assecondare il modello di annotazione basato su frames e, più precisamente, può essere impiegato per aggiungere un secondo strato, tipicamente semantico, di annotazione a testi già analizzati a livello sintattico. SALTO presenta un'interfaccia intuitiva, grazie a un editor visuale che rende agevole l'annotazione.

Dopo aver acquisito la competenza dell'uso del software, è stata presentata una lista di 41 frames diversi e una tabella (tab. 1, pag. 41), in cui erano stati appuntati, per ciascun frame, i target individuati e il loro numero complessivo. È stato quindi richiesto di trovare, per ogni situazione semantica, nuove istanze di target che la evocassero. In questa circostanza, sono state coinvolti nella ricerca anche target non verbali, assumendo che una frase potesse contenere perfino più di un'unità lessicale evocatrice. Nella scelta dei target idonei, sono stati impiegati vari criteri: si sono ricercate forme identiche o differenti rispetto a quelle già registrate, e sinonimi, ricorrendo talvolta al thesaurus di FrameNet. Dopo aver individuato un numero sufficiente di frasi, queste sono state annotate mediante l'uso di SALTO.

Entrambe le fasi del lavoro di annotazione si sono concluse con risultati soddisfacenti, nonostante non siano mancati aspetti problematici su cui riflettere. La difficoltà principale del primo task si è rivelata nella ricerca delle dieci frasi più rappresentative per un frame. È stato necessario rispettare una serie di regole nella selezione delle frasi: anzitutto, la ricerca si è basata esclusivamente su target

di tipo verbale, escludendo dal novero parole con funzioni sintattiche diverse dal predicato. Per esempio, date le frasi:

> Questa relazione (…) stabilisce di dare [Theme degli aiuti] [Recipient a quegli Stati che hanno bisogno di denaro per i loro bilanci].
>
> (…) trattandosi di una donazione [Manner volontaria].

la prima frase è considerata buona, poiché il frame *Giving* è evocato dal target verbale *dare*. La seconda non soddisfa i requisiti richiesti dalla selezione, in quanto il target si presenta in forma sostantivale.

Inoltre, è stato richiesto di prediligere verbi diversi e costrutti sintattici differenti, al fine di fornire una visione quanto più ampia e articolata possibile della variabilità dei target. È il caso delle seguenti frasi, che evocano il frame *Telling* con target diversi:

> (…) che confermi l'impegno a informare [Addressee il Parlamento] [Topic di misure di assistenza eccezionali] nel momento in cui vengono adottate.
>
> [Addressee Mi] dicono [Message che il riferimento alla Macedonia è stato comunque di fatto cancellato].

Infine, l'ultimo criterio adottato è stato quello di evitare frasi che includessero il target in una subordinata relativa; questo perché il calcolatore tende ad annotare il pronome relativo come parte del target e i termini antecedenti al relativo come istanze dei frame elements. Di seguito, un caso del frame *Death* a titolo esemplificativo, così come è stato pre-annotato dal sistema automatico:

> In Europa, oggigiorno, vediamo ancora [Protagonist animali soggetti a sofferenze terribili ed animali] che muoiono in transito a [Cause causa delle condizioni disumane di trasporto].

L'annotazione è errata, in quanto il pronome *che* non può far parte del target (*muoiono*), bensì istanzia i ruoli semantici a cui si riferisce, espressi nella proposizione principale (*animali*). La selezione delle dieci frasi è quindi avvenuta solo in base a target correttamente individuati, tralasciando situazioni del tipo

appena descritto. Un esempio di corretta individuazione del target per il frame *Death* è il seguente:

> [Time Alcuni giorni dopo] (…) moriva [Manner improvvisamente] [Protagonist il Presidente Rugova].

Per facilitare la fase preliminare di addestramento, l'indagine si è concentrata prima su frames per la descrizione di situazioni concrete (ad esempio *Killing*, *Death*, *Giving*), in modo da familiarizzare con il riconoscimento delle strutture semantiche più semplici. In seguito, la ricerca ha coinvolto anche i frames più complessi, riferiti a situazioni semantiche astratte. Un aspetto problematico della ricerca delle dieci frasi è la distinzione tra frames molto simili: alcune strutture semantiche presentano leggere variazioni di significato rispetto ad altre, al punto tale da veicolare delle ambiguità nella scelta delle frasi appropriate. Bisogna quindi prestare estrema attenzione al senso della frase, per evitare di assegnare il frame inadeguato. La questione emerge, per esempio, nei frames *Perception_active* e *Perception_experience*: nel primo frame, un soggetto rivolge intenzionalmente la propria attenzione a un'entità o un fenomeno al fine di avere un'esperienza percettiva; nel secondo, il soggetto subisce un'esperienza percettiva in maniera involontaria. Sebbene i verbi che evocano queste situazioni possano essere gli stessi (*vedere*, *sentire*, *gustare*), il senso della frase cambia in base all'intenzionalità dell'esperienza percettiva. Altri casi di affinità sono stati riscontrati nei frames:

1. *Awareness* e *Cogitation*;
2. *Discussion* e *Speak_on_topic*;
3. *Evidence* e *Reasoning*;
4. *Questioning* e *Request*;
5. *Speak_on_topic*, *Statement* e *Telling*.

Oltre a distinguere le differenze tra strutture semantiche simili, è stato necessario anche disambiguare il senso di alcune parole che avrebbero potuto evocare più di un frame. Un esempio è dato dal verbo *provare*, che può richiamare sia il frame *Feeling* che il frame *Attempt*:

(…) quello spirito di democrazia ed equilibrio [Emotion che] già abbiamo potuto provare nella passata legislatura.

(…) ma ci abbiamo provato e ci siamo riusciti in più occasioni.

Mentre nella prima frase il verbo indica un'*esperienza percettiva*, nella seconda il medesimo verbo assume un senso diverso, veicolando il significato di *tentare*. Un'ultima questione rilevante consiste nella definizione del *semantic type* per alcune unità lessicali; ovvero, le parole incorporano nel loro significato un tipo semantico, che in alcuni casi assume valore positivo o negativo. Considerando, ad esempio, che il frame *Memory* è evocato tanto dal verbo *ricordare* quanto dal verbo *dimenticare*, emerge che il semantic type della prima parola ha un valore positivo, mentre nella seconda ha un valore negativo. Pertanto, alcuni frames hanno la capacità di evocare una situazione semantica e il suo contrario. Fanno parte del novero i frames *Certainty*, *Compliance*, *Judgment_Communication*, *Memory* e *Possession*. Si vedano le frasi scelte per il frame *Compliance*:

> Ma che cosa accade quando [Protagonist un appaltatore] infrange [Norm le regole]?
>
> [Norm I diritti fondamentali dei bambini] verranno rispettati.

Altri esempi significativi sono forniti dal frame *Judgment_Communication*:

> (…) [Communicator la onorevole Green] ha appena lodato [Evaluee il Presidente].
>
> (…) [Communicator il primo ministro Yilmaz, il vice primo ministro e numerosi ministri e deputati] hanno condannato [Evaluee questo attentato].

Dopo aver definito le frasi rappresentative per ogni struttura semantica, si è svolta la fase di revisione e correzione dei frame elements. La difficoltà principale è sorta nella definizione dei confini dei ruoli semantici: ossia in quali casi includere la punteggiatura e se comprendere i verbi ausiliari nel target. Entrambi i dubbi si sono risolti grazie alla consultazione del corpus di addestramento, il quale mostra che i segni di interpunzione esterni al sintagma non rientrano nel frame element e che i verbi ausiliari non fanno parte del target.

Nel secondo task del progetto, mirato ad arricchire la ISST con ulteriori frasi annotate, l'aspetto più impegnativo si è presentato nella ricerca di nuove istanze di frames da aggiungere alla collezione. L'indagine si è svolta su un campione di 3505 frasi: con un documento di tali dimensioni, scorrere l'intera collezione analizzando le frasi una ad una sarebbe stato estremamente costoso in termini di tempo di elaborazione. Per questa ragione, il lavoro di indagine ha avanzato per tentativi, tramite la definizione di una serie di possibili target e la loro ricerca nel documento. Data una tabella con i frames e i target già rilevati, sono state cercate nuove unità lessicali che evocassero quei frames: la ricerca è avvenuta a partire dallo stesso lemma, da forme diverse derivate dal lemma o da sinonimi rispetto ai target individuati in precedenza. Mentre nella prima parte dello stage l'annotazione ha coinvolto solo target di tipo verbale, in questa fase l'analisi si è estesa anche ai sostantivi: esemplari i casi del frame *Buldings*, evocato da nomi come *casa* o *edificio*, e del frame *Cause_to_end*, richiamato dal termine *cancellazione*. In seguito al rilevamento dell'occorrenza di un frame in una frase, la tabella è stata aggiornata in modo da riportare il numero corrente di istanze individuate. Successivamente, ciascuna frase è stata annotata con il software SALTO, ovvero recuperando la frase dalla collezione mediante il suo identificatore univoco e marcando correttamente i frame elements. In fig. 1, pag. 42, un'immagine del formato di annotazione di SALTO.

Confrontando le due fasi del progetto, la prima ha certamente richiesto un lasso di tempo maggiore per portare a termine il compito. Le cause principali di questa necessità sono state due: anzitutto, il periodo di apprendimento che ha preceduto la prima esperienza di annotazione, con l'introduzione dei metodi e delle nozioni essenziali; successivamente, la cautela profusa nella scelta delle dieci frasi opportune per ogni frame. La seconda fase del lavoro ha invece mantenuto un andamento notevolmente più deciso e spedito, grazie alle conoscenze precedentemente acquisite e poi consolidate dalla pratica nell'annotazione. Le difficoltà comuni ai due task si sono verificate nella correzione e nell'assegnazione dei ruoli semantici, in quanto l'individuazione dei frame elements ha caratterizzato entrambi gli stadi del progetto. L'impiego di SALTO durante la seconda fase ha inoltre comportato un consistente miglioramento nella rapidità dell'annotazione. Sebbene il progetto abbia presentato una serie di aspetti

tanto interessanti quanto problematici, ogni perplessità in merito all'annotazione si è risolta tramite il ricorso al corpus di addestramento, un solido riferimento da consultare nei casi di incertezza.

5. Conclusioni

La relazione ha presentato gli aspetti salienti del processo di valutazione, con la premessa che la scelta di un metodo valutativo adeguato al sistema dipende da alcuni fattori: tale scelta è anzitutto determinata dal ciclo di vita del sistema, ovvero dallo stadio di crescita in cui la tecnologia si trova. Inoltre, bisogna considerare il tipo di input e output relativi al sistema, in quanto ogni strumento necessita di precise metriche valutative. Si è visto che nei sistemi di analisi la valutazione può impiegare un *gold standard* di riferimento, basarsi sulle caratteristiche del sistema o ricorrere al metodo del componente incorporato. Le tecniche cambiano se si considera un sistema per la produzione di output linguistici, per cui vengono introdotte le misure intrinseche ed estrinseche. Ancora, la metodologia valutativa muta nel caso di sistemi interattivi, dove l'attenzione si rivolge al reciproco rapporto tra utente e computer. Prescindendo dalle differenti forme assunte dalla valutazione, è ragionevole osservare i vantaggi di un tale approccio: il costante interesse verso le prestazioni dei sistemi di NLP incoraggia la creazione di prodotti idonei alle esigenze degli utenti, oltre che coerenti agli scopi per cui sono stati progettati; in più, le sfide proposte dal progresso tecnologico promuovono la definizione di task sempre più complessi e interessanti; infine, non bisogna trascurare l'impatto propriamente sociale dell'attività di valutazione, che da un lato offre un'opportunità di collaborazione tra linguisti e informatici, dall'altro dischiude uno spazio condiviso in cui gli sviluppatori dei sistemi possono confrontarsi. Per queste ragioni, la nascita di Evalita ha comportato notevole progresso nell'applicazione dei metodi valutativi agli strumenti per la lingua italiana. Le innovazioni si sono verificate già a partire dall'edizione 2009, in cui sono state definite misure per la valutazione di *speech tasks*. L'edizione 2011 ha poi introdotto la valutazione di sistemi per l'annotazione semantica, questione estremamente attuale e interessante, poiché permette di implementare tasks basati sulla conoscenza del mondo e sulla costruzione di ontologie: è il caso di *Question Answering*, *Information Extraction* e *Information Retrieval*, fondati sui compiti di Frame Labeling e Super Sense Tagging.

La partecipazione alla realizzazione del progetto per FLaIT ha rivelato una valutazione abbastanza positiva dei sistemi di Frame Labeling, i quali hanno ottenuto dei punteggi di precisione e recall piuttosto alti. Il sottotask eseguito in maniera più adeguata è stato il Frame Labeling, seguito dai risultati apprezzabili di Argument Classification, e da quelli meno precisi di Boundary Detection. In generale, i sistemi hanno eseguito agevolmente il task sia grazie al numero esiguo di frames da riconoscere, sia per la presenza di target già marcati nel corpus di prova. L'esperienza legata al progetto ha permesso uno stretto contatto con gli strumenti e le metodologie di annotazione impiegate in Evalita, evidenziando le doti e le complessità dell'annotazione semi-automatica. Tuttavia, la questione relativa all'annotazione semantica è notevolmente recente: lo sviluppo di sistemi di Frame Labeling richiede ulteriori e approfondite ricerche sul campo, al fine di annotare in maniera adeguata i testi con informazione semantica.

6. Appendice

Frame	N° di istanze	LUs (forme)						
Request	9	chiede	chiedere	invita	sollecitando-	chiesto	ordina	ordinò
Sending	9	inviata	mandare	mandato	spedito	restituito	inviare	rimandarono
State_of_entity	8	è	era	sono	sarebbero	stare	stiamo	
Being_located	7	situati	è	era	situato	Situato		
Competition	7	andata	giocasse ro	giocare	giocano	gioca	gare	giocato
Congregating	7	incontrato	trova	Incontrar-	Incontrano	incontrassi	incontra	
Communication	6	parlato	parla	parlare	rivolge	parlano	sottolineato	
Escaping	6	fuga	scappa	scappano	abbandonano	fuggiti	fuggire	
Event	6	avvenuta	svolto	avvenuto	successo	capita		
Activity_prepare	5	preparando	preparare	prepara				
Cause_to_start	5	provocò	provocate	provocare	provoca			
Communication_response	5	risponde	rispondevo	ribatte				
Motion_directional	4	scende	Tuffar-	innalzato	accascia			
Quitting_a_place	4	abbandonano	abbandonare	lasciato				
Risolve_problem	4	risolvere	riequilibrare	risolvere				
State_continue	4	resta	stare	stabilizzato	rimase			
Using	4	serviva	impiegati	usavano	usa			
Wearing	4	indossavano	indossa	veste				
Referring_by_name	3	chiamati	chiamavano	chiamava				
Removing	3	eliminare	scacciare	sgombrare				
Respond_to_proposal	3	scartata	accettato					
Scrutiny	3	scrutare	Studiando	analizziamo				
Speak_on_topic	3	parla	parlare	parleremo				
Subjective_influence	3	spinge	spingono	spinto				
Undergo_change	3	cambiato	cambia					

Unknown	3	scivola	raggiunto	lascia				
Use_firearm	3	sparano	Sparano	sparato				
Visiting	3	visitato	Visitar-					
Accomplishment	2	compiuto	compirà					
Activity_stop	2	abbandonato	finito					
Assessing	2	valutate	valutò					
Avoiding	2	evita	Evita					
Being_obligated	2	costretta						
Buildings	2	casa						
Cause_to_end	2	cancellazione	finirà					
Cause_to_make_noise	2	suona						
Cause_to_make_progress	2	svolto	sviluppa					
Ceasing_to_be	2	spariti	sparire					
Change_position_on a_scale	2	salire	scenderà					
Choosing	2	scelto	scegli					
Commerce_pay	2	pagando	pagare					

Tabella 1: per ciascun frame sono segnati i target individuati nel testo e il loro numero complessivo.

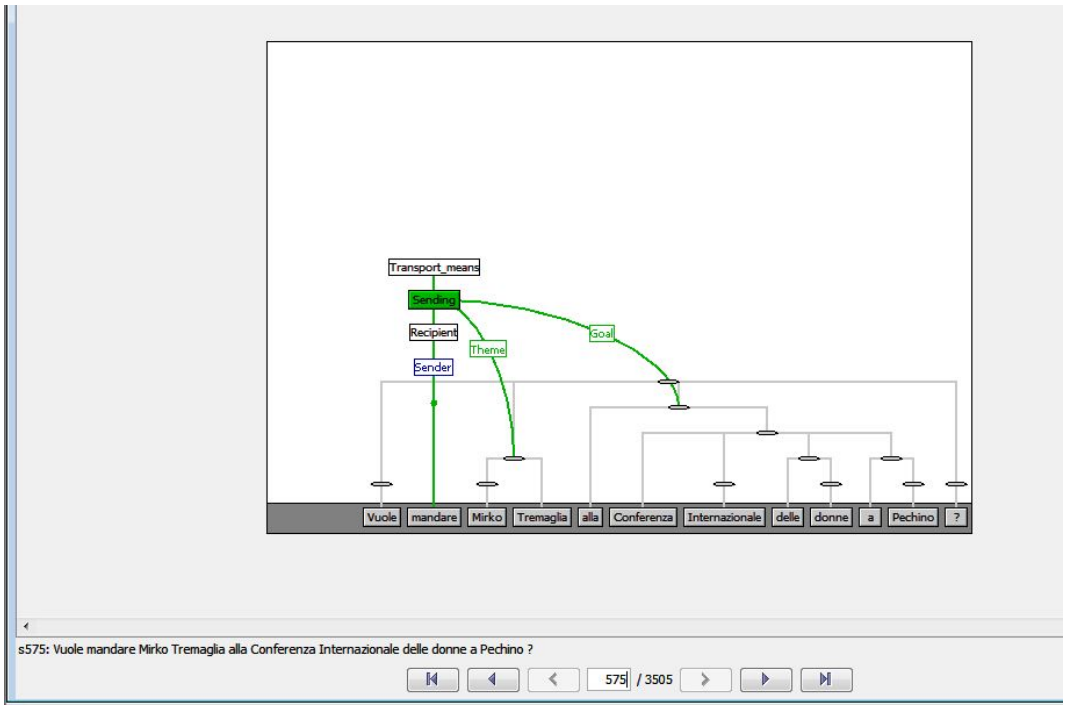

Figura 1: formato di annotazione di SALTO.

7. Bibliografia

Atkins, Sue, Michael Rundell e Hiroaki Sato. 2003. *The contribution of FrameNet to practical lexicography*. In: "International Journal of Lexicography", 3, pp. 333-357.

Baker, C.F., Fillmore, C.J., Lowe, J.B. 1998. *The Berkeley FrameNet Project*. In: Proceedings of the 36th ACL Meeting and 17th ICCL Conference, Morgan Kaufmann.

Basili, R., De Cao, D., Croce, D., Coppola, B., Moschitti, A. 2009. *Cross-language frame semantics transfer in bilingual corpora*. In: Proc. of 10th Int. Conf. On Intelligent Text Processing and Computational Linguistics (CICLing 2009), Mexico City, Mexico.

Basili, R., De Cao, D., Lenci, A., Moschitti, A., and Venturi, G.: Evalita 2011: *the Frame Labeling over Italian Texts Task*. In: Working Notes of EVALITA 2011, 23-24th January 2012, Rome, Italy, ISSN 2240-5186 (2012).

Cutugno, F. and Falcone, M.: *Evalita speech tasks*. In: Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, 12th December 2009, Reggio Emilia, Italy, ISBN 978-88-903581-1-1 (2009).

Dei Rossi, S., Di Pietro, G., and Simi, M.: Evalita 2011: *Description and Results of the SuperSense Tagging Task*. In Working Notes of EVALITA 2011, 23-24th January 2012, Rome, Italy, ISSN 2240-5186 (2012).

EAGLES. 1996. *EAGLES: evaluation of natural language processing systems.* Final Report.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Fillmore, Charles J., Christopher R. Johnson e Miriam R.L. Petruck. 2003. *Background to FrameNet*. In: "International Journal of Lexicography", 3, pp. 235-250.

Fillmore, Charles J., Miriam R.L. Petruck, Josef Ruppenhofer e Abby Wright. 2003. *FrameNet in action: the case of attaching*. In: "International Journal of Lexicography", 3, pp. 297-332.

Hirschman, Lynette, e Inderjeet Mani. 2003. *Evaluation*. In: Ruslan Mitkov. *The Oxford handbook of computational linguistics*. New York, Oxford University Press, pp. 414-429.

Jarke, M., J. Turner, E. Stohr, Y. Vassiliou, N. White e K. Michielson. 1985. *Field evaluation of natural language for data retrieval.* In: IEEE Transactions on Software Engineering, pp. 97-133.

Lenci, A., S. Montemagni, E. Vecchi, G. Venturi. *Enriching the isst-tanl corpus*

with semantic frames. In: Forthcoming.

Magnini, Bernardo, e Amedeo Cappelli. 2007. *Evalita 2007: evaluating natural language tools for Italian*. "Intelligenza Artificiale", 2, p. 3.

Magnini, B. and Cappelli, A.: *Introduction to Evalita 2009*. In: Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, 12th December 2009, Reggio Emilia, Italy, ISBN 978-88-903581-1-1 (2009).

Mani, I. e E. Bloedorn. 1998. *Machine learning of generic and user-focused summarization.*In: Proceedings of AAAI '98, Madison, Wis.

Montemagni, S., F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, Saracino, F. Zanzotto, N. Mana, F. Pianesi, R. Delmonte. 2003. *Building the Italian Syntactic–Semantic Treebank.* In: Anne Abeill´e (ed.), "Building and Using syntactically annotated corpora", Kluwer, Dordrecht.

Montemagni, S. et al. 2003. *Building the Italian Syntactic-Semantic Treebank*. In: Building and using Parsed Corpora, Language and Speech series, pp. 189–210.

Palmer, M., Gildea, D., Kingsbury, P. 2005. *The Proposition Bank: A Corpus Annotated with Semantic Roles*. In: Computational Linguistics Journal, 31(1).

Polifroni, J., S. Seneff, J. Glass e T. Hazen. 1998. *Evaluation methodology for a telephone-based conversational system.* In: Proceedings of the 1st International Conference on Language Resources and Evaluation, Granada, pp. 43-50.

Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, 12th December 2009, Reggio Emilia, Italy, ISBN 978-88-903581-1-1 (2009).

Robin, J. 1994. *Revision-based generation of natural language summaries providing historical background: corpus-based analysis, design and implementation.* Ph.D. thesis, Columbia University.

Sito web di Evalita, pagine *Evalita 2007*, *Evalita 2009*, *Evalita 2011*
http://www.evalita.it/2007

http://www.evalita.it/2009

http://www.evalita.it/2011.

Sito web di FrameNet, pagina *About FrameNet*

https://framenet.icsi.berkeley.edu/fndrupal/about.

Sito web di Ontotext, pagina *Semantic Annotation*

http://www.ontotext.com/kim/semantic-annotation.

Sito web di SALTO, pagine *Home* e *Documentation*

http://www.coli.uni-saarland.de/projects/salsa/page.php?id=index.

http://www.coli.uni-saarland.de/projects/salsa/salto/doc/.

Sito web di WordNet, pagina *What is WordNet?*

http://wordnet.princeton.edu/

Tonelli, S., E. Pianta. 2008. *Frame information transfer from english to italian*. In: Proc. of LREC Conference, Marrakech, Marocco.

Walker, M., D. Litman, C. Kamm e A. Abella. 1998. *Evaluating spoken dialogue agents with PARADISE: two cases studies.* In: Computer Speech and Language, 12(4).

Whittaker, S. e M. Walker. 1989. *Comparing two user-oriented database query languages: a field study.* Technical Report HPL-ISC-89060, Hewlett Packard Laboratories, Bristol.

Working Notes of EVALITA 2011, 23-24th January 2012, Rome, Italy, ISSN 2240-5186 (2012).

Wikipedia, voce *Natural Language Processing*

http://en.wikipedia.org/wiki/Natural_language_processing.