Passo 2
Traduzione dei lemmi

Traduzione dei lemmi (creazione dei Babel synset)

Fino ad ora siamo riusciti ad ottenere μ, ossia a collegare le Wikipages inglesi ai sensi di WordNet (ovviamente inglesi).

Ora, data una Wikipage w e il suo mapping μ(w) = s, creiamo un Babel synset S ∪ W dove S è il synset di WordNet a cui s appartiene e W include:

w

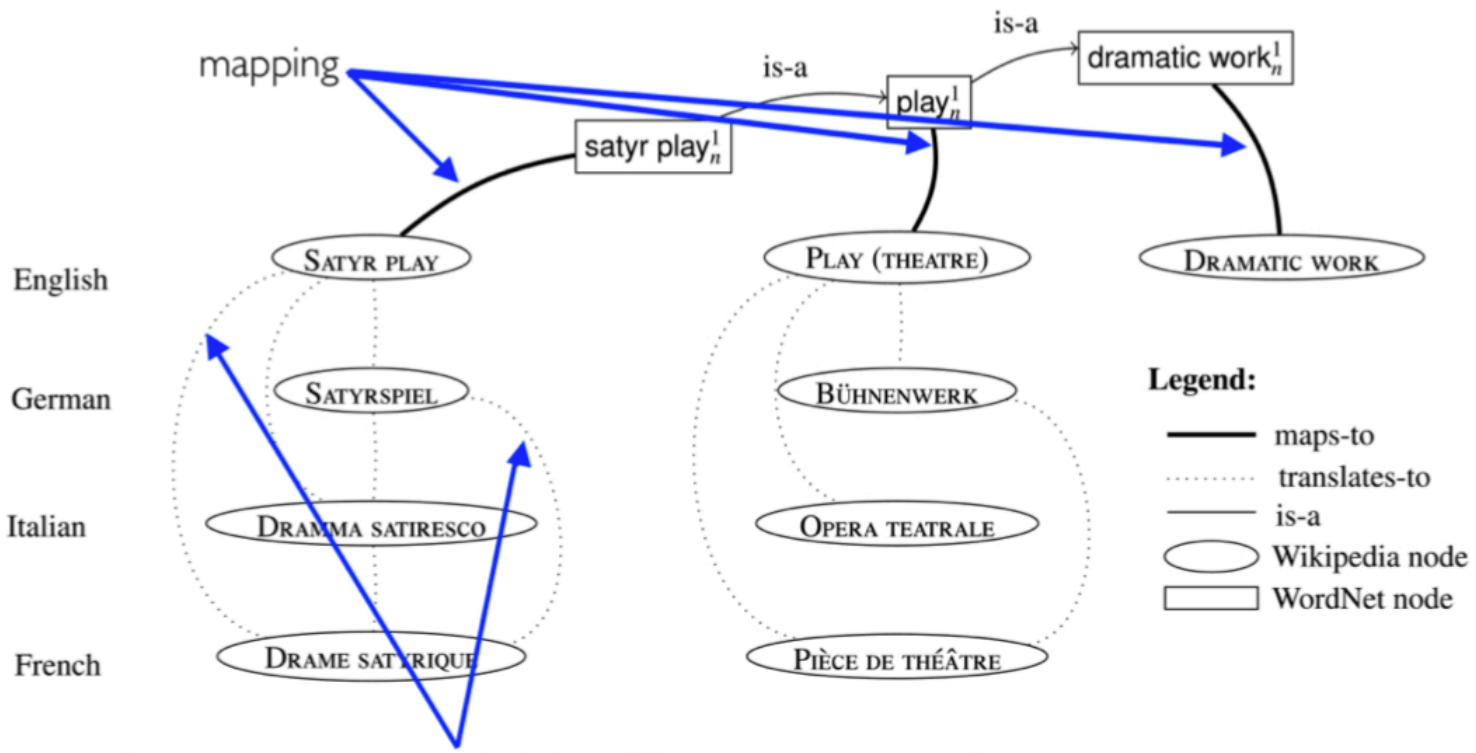
Il set di titoli delle Wikipages che linkano a w.

Tutte le pagine linkate da w mediante link inter-lingua (ossia la traduzione della Wikipage in altre lingue - ricordiamo che stiamo trattando solo il titolo e non la pagina per intero -).

Le redirezioni ai link inter-lingua trovati nella Wikipedia del linguaggio target.

Babel synsets

Comprendiamo ora quanto avevamo anticipato poco fa: i Babel synset contengono tutti i sinonimi presi da WordNet per un dato concetto, con l'aggiunta di tutti i lemmi di altre lingue che rappresentano il concetto espresso dal Babel synset.



Babel synsets integrating WordNet synsets and Wikipages are straightforwardly translated by collecting the hyperlinks to wikipedias in languages other than English

vediamo una porzione di BabelNet relativa al concetto play.

Problemi

Possiamo incontrare due diverse problematiche durante l'esecuzione di questa fase di creazione dei babel synsets:

1. Un concetto potrebbe essere coperto soltanto su WordNet oppure su Wikipedia, ossia, non è stato trovato un mapping per w oppure un senso s non è stato mappato su nessun w. In questi casi, non è possibile stabilire alcun link.

2. Anche se un concetto è coperto sia da WordNet che da Wikipedia, potrebbe non esserci una traduzione in ogni linguaggio di interesse per quel concetto (*es.* i link inter-language per Spagnolo e Catalano della pagina *Play(theatre)* non sono presenti in Wikipedia).

Corpus esterno

Usare un corpus esterno

Una metodologia utile per tradurre i concetti rappresentati dai Babel synset aumentando la copertura (ovvero, non limitandosi all'uso di inter-language links) consiste nello sfruttare un corpus esterno, il **`SemCorr`**, con contiene più di 200.000 parole annotate con sensi di WordNet.

In sostanza, disponendo di SemCor e supponendo di dover costruire il Babel synset di Play(theatre), procediamo come segue:

1. Consideriamo innanzi tutto il senso WordNet associato alla pagina Play(theatre), che supponiamo essere $play^{1}_{n}$

2. Cerchiamo tutte le occorrenze di $play^{1}_{n}$ in SemCor e memorizziamo in una collezione C tutte le parole associate a quel senso.

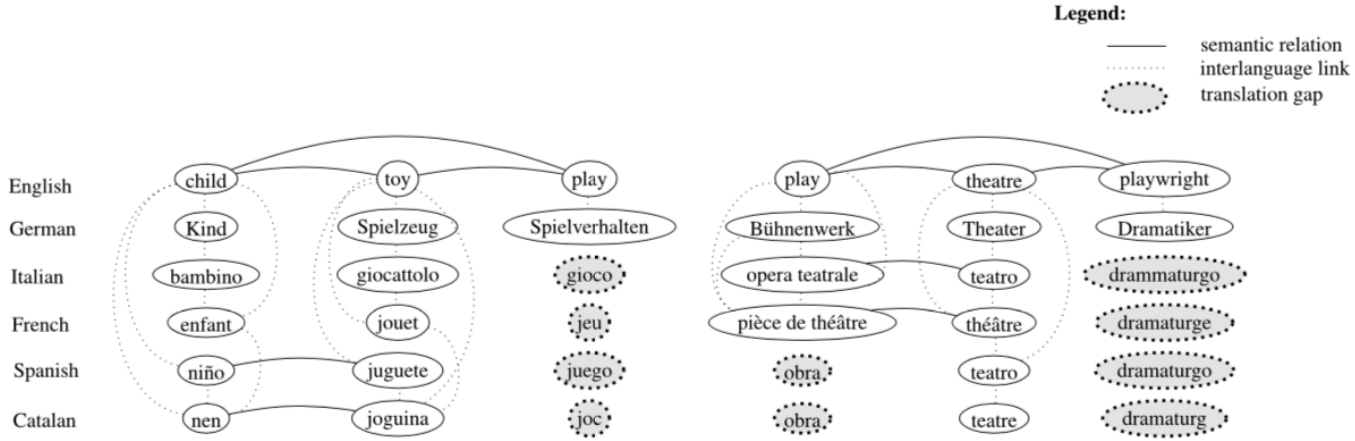
3. Prendiamo poi ogni Wikipage w che linka a *Play(theatre)* e inseriamo in C le frasi in cui è presente l'hyperlink verso Play (theatre) (debitamente ripulite dalle stop words).

Corpus esterno

A questo punto C sarà costituito da un'insieme di parole che sono in qualche modo legate a $play^1_n$

Andiamo a ripetere il secondo punto per tutti gli altri elementi inglesi presenti nello stesso Babel synset di $play^1_n$ (ricordiamo che il synset è già stato creato e stiamo solo cercando di ampliarlo con lessicalizzazioni multilingua) al fine di arricchire ulteriormente C.

Non resta che adoperare un sistema di machine translation per tradurre tutte le parole contenute in C nelle varie lingue di interesse e, una volta identificata la traduzione più probabile/frequente per ciascuna delle parole, la andiamo ad aggiungere al Babel synset. Si noti che la traduzione farà anche uso del contesto al fine di migliorare le sue performance.



Il Babel synset iniziale è stato esteso con le parole $drame_{fr}$, $dramma_{it}$, $obra_{ca}$, $obra_{es}$. Le lessicalizzazioni aggiunte sono indicate da ellissi tratteggiati