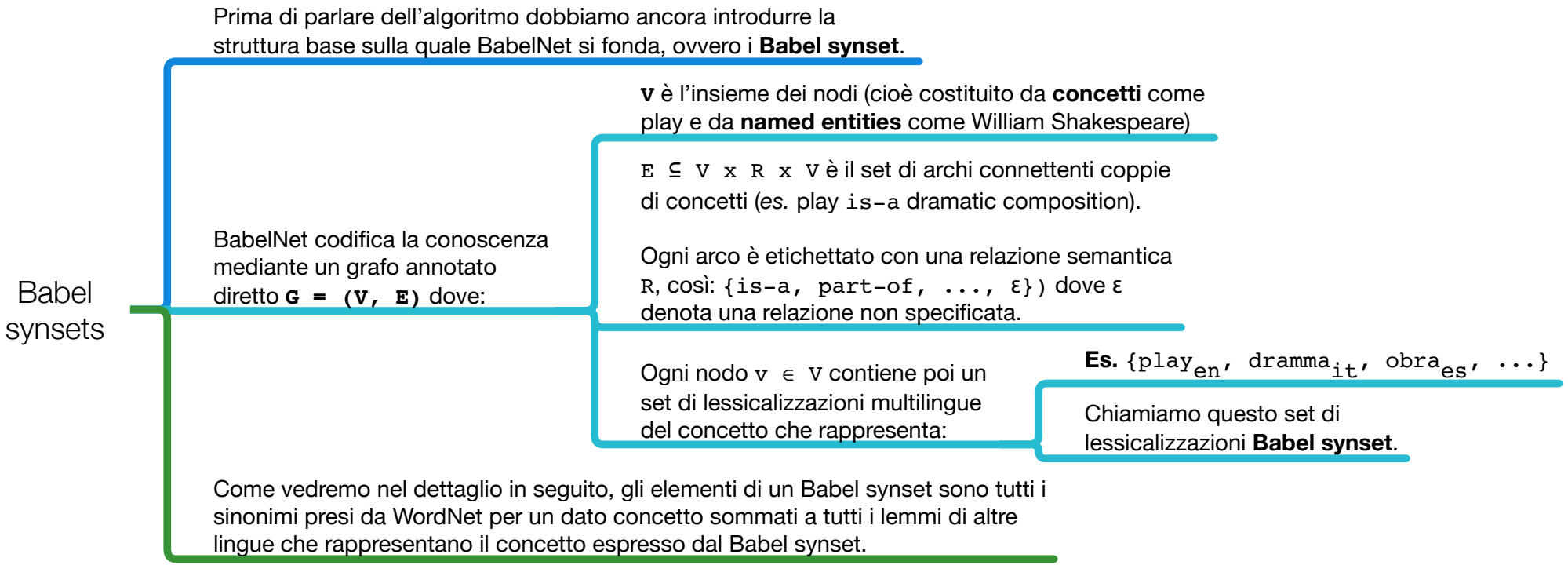
Fusione Wiki + WordNet

Andiamo finalmente a studiare l'algoritmo mediante il quale Wikipedia e WordNet vengono automaticamente fuse per costruire BabelNet.


Babel synsets
Prima di parlare dell'algoritmo dobbiamo ancora introdurre la struttura base sulla quale BabelNet si fonda, ovvero i **Babel synset**.
BabelNet codifica la conoscenza mediante un grafo annotato diretto **G = (V, E)** dove:
V è l'insieme dei nodi (cioè costituito da **concetti** come play e da **named entities** come William Shakespeare)
E ⊆ V x R x V è il set di archi connettenti coppie di concetti (*es.* play is-a dramatic composition).
Ogni arco è etichettato con una relazione semantica R, così: {is-a, part-of, ..., ε}) dove ε denota una relazione non specificata.
Ogni nodo v ∈ V contiene poi un set di lessicalizzazioni multilingue del concetto che rappresenta:
Es. {play$_{en}$, dramma$_{it}$, obra$_{es}$, ...}
Chiamiamo questo set di lessicalizzazioni **Babel synset**.
Come vedremo nel dettaglio in seguito, gli elementi di un Babel synset sono tutti i sinonimi presi da WordNet per un dato concetto sommati a tutti i lemmi di altre lingue che rappresentano il concetto espresso dal Babel synset.

Costruire il grafo

Per costruire il grafo di BabelNet vengono prima di tutte raccolte le seguenti informazioni:

Da WordNet, tutti i significati (come concetti) e tutte le relazioni lessicali e semantiche fra synsets (come relazioni etichettate a seconda della tipologia che avevano in WordNet).

Da Wikipedia, tutte le Wikipage (come concetti) e tutti i loro internal link (come relazioni semanticamente non specificate). Vengono anche recuperati i redirection links e le categories di ogni Wikipage (sempre come relazioni semanticamente non specificate).

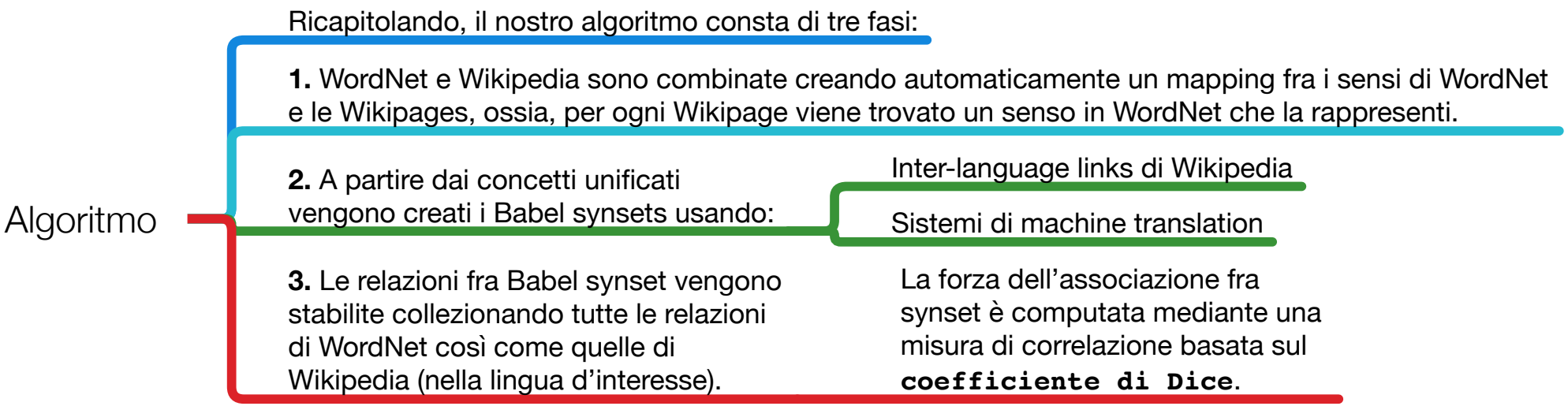
Costruire il grafo

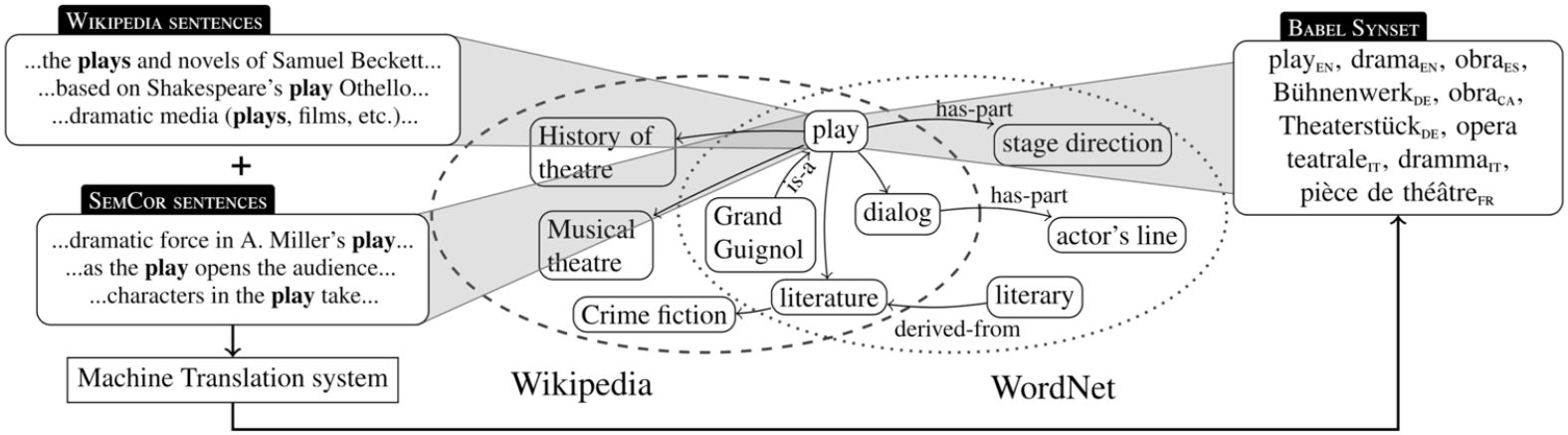
Giacché i concetti recuperati da Wikipedia e da WordNet possono sovrapporsi sia in termini di concetti che in termini di relazioni è necessario effettuare effettuare una unificazione dei dati rimuovendo i doppioni.

In altre parole, partendo dall'assunzione che Wikipedia sia una risorsa più estesa ma meno formalizzata di WordNet, cerchiamo di importare la struttura solida di WordNet all'interno di Wikipedia.

Una volta fatto questo, per poter implementare le funzionalità multilingua è necessario recuperare le realizzazioni lessicali in altre lingue di tutti i concetti che abbiamo recuperato (vedremo dopo come farlo) al fine di creare dei Babel synsets.

Infine, i vari synset sono collegati fra loro mediante quelle relazioni semantiche già recuperate da WordNet e da Wikipedia.


Algoritmo
Ricapitolando, il nostro algoritmo consta di tre fasi:
1. WordNet e Wikipedia sono combinate creando automaticamente un mapping fra i sensi di WordNet e le Wikipages, ossia, per ogni Wikipage viene trovato un senso in WordNet che la rappresenti.
2. A partire dai concetti unificati vengono creati i Babel synsets usando:
Inter-language links di Wikipedia
Sistemi di machine translation
3. Le relazioni fra Babel synset vengono stabilite collezionando tutte le relazioni di WordNet così come quelle di Wikipedia (nella lingua d'interesse).
La forza dell'associazione fra synset è computata mediante una misura di correlazione basata sul coefficiente di Dice.



La Figura illustra un piccolo estratto di BabelNet esplicitando la provenienza dei nodi (concetti)

BabelNet
figura

Nota:
Gli archi senza etichetta sono ottenuti dagli internal link delle WikiPages mentre quelli etichettati provengono da WordNet