Enriched Genic Interaction Extraction Challenge Data Format

1 File Structure

LLL challenge training data and the linguistic information are represented as follows. The file consists of the following fields (one field by line) :

- ID : unique identifier of the Pubmed abstract that contains the sentence and the sentence position number
- sentence : the original sentence
- words : sequence of the sentence words
- agents : list of the agents of the genic interactions
- targets : list of the targets of the genic interactions
- genic_interactions : list of the interactions described in the sentence.
- lemmas : list of identified canonical form of words
- syntactic_relations : list of the syntactic relations in the sentence. See the Syntactic Analysis Guidelines for more information about this field

2 Field Structure

A tab separates each element of a field :

ID

The ID field contains the abstract PubMed ID (PMID) which the sentence is extracted from and it contains the sentence number in this abstract.

ID	(tabulation)	11011148-1

SENTENCE

This field contains the sentence.

sentence	(tabulation)	ykuD was transcribed by SigK RNA polymerase from T4 of sporulation.

WORDS, AGENTS, TARGETS, GENIC_INTERACTIONS, LEMMAS, SYNTACTIC_RELATIONS

Other fields are organised according to the following format :

Field_Name	(tabulation)	predicate1(argument1_1,argument1_2,...)
	(tabulation)	predicate2(argument2_1,argument2_2,...)
	(tabulation)	...

EXAMPLE WORDS

words	word(0,'ykuD',0,3) word(1,'was',5,7) word(2,'transcribed',9,19) word(3,'by',21,22) word(4,'SigK',24,27) word(5,'RNA',29,31) word(6,'polymerase',33,42) word(7,'from',44,47) word(8,'T4',49,50) word(9,'of',52,53) word(10,'sporulation',55,65)

3 Predicate Description

WORD

The predicate "word" refers to a word of the sentence and accepts four arguments :
word(id_word,'string_word',start_word,end_word)

id_word	integer, unique word id
string_word	string, the actual word
start_word	integer, position of the first character in the sentence (starting at 0)
end_word	integer, position of the last character in the sentence (starting at 0)

AGENT

The predicate "agent" refers to the agent of the genic interaction. It accepts one argument :
agent(id_word)

id_word	integer, id of the word the agent refers to

TARGET

The predicate "target" refers to the target of the genic interaction. It accepts one argument :
target(id_word)

id_word	integer, id of the word the target refers to

GENIC_INTERACTION

The predicate "genic_interaction" refers to an interaction between an agent and a target :
genic_interaction(id_word1,id_word2)

id_word1	integer, id of the word the agent refers to
id_word2	integer, id of the word the target refers to

LEMMAS

The predicate "lemma" refers to the normalized form (lemma) of a word.
lemma(id_word,'string_lemma')

id_word	integer, id of the word the lemma refers to
string_lemma	string,the lemma of the word

SYNTACTIC_RELATION

The predicate "relation" refers to the normalized form (lemma) of a word. See the Syntactic Analysis Guidelines for more information.
relation('string_relation',id_word1,id_word2)

string_relation	string, the information contained in a syntactic relation (function of the relation :morpho-syntactic nature of the 2 words)
id_word1	integer, id of the first word (the head) linked by the relation
id_word2	integer, id of the second word (the expension) linked by the relation

4 Example

ID	10747015-5
sentence	Localization of SpoIIE was shown to be dependent on the essential cell division protein FtsZ.
words	word(0,'Localization',0,11) word(1,'of',13,14) word(2,'SpoIIE',16,21) word(3,'was',23,25) word(4,'shown',27,31) word(5,'to',33,34) word(6,'be',36,37) word(7,'dependent',39,47) word(8,'on',49,50) word(9,'the',52,54) word(10,'essential',56,64) word(11,'cell',66,69) word(12,'division',71,78) word(13,'protein',80,86) word(14,'FtsZ',88,91)
lemmas	lemma(0,'localization') lemma(1,'of') lemma(2,'spoIIE') lemma(3,'be') lemma(4,'show') lemma(5,'to') lemma(6,'be') lemma(7,'dependent') lemma(8,'on') lemma(9,'the') lemma(10,'essential') lemma(11,'cell') lemma(12,'division') lemma(13,'protein') lemma(14,'ftsZ')
syntactic_relations	relation('comp_of :N-N',0,2) relation('mod_att :N-ADJ',13,10) relation('mod_pred :N-ADJ',0,7) relation('mod_att :N-N',14,13) relation('mod_att :N-N',12,11) relation('mod_att :N-N',13,12) relation('comp_on :ADJ-N',7,14)
agents	agent(14)
targets	target(2)
genic_interactions	genic_interaction(14,2)