Performance and limitations of the linguistically motivated Cocoa/Peaberry system in a broad biomedical domain.

S. V. Ramanan
RelAgent Private Ltd.
56, Venkatratnam Nagar
Adyar, Chennai 600020
ramanan@npjoint.com

P. Senthil Nathan
RelAgent Private Ltd.
56, Venkatratnam Nagar
Adyar, Chennai 600020
senthil@npjoint.com

Abstract

We tested a linguistically motivated rule-based system in the Cancer Genetics task of the BioNLP13 shared task challenge. The performance of the system was very moderate, ranging from 52% against the development set to 45% against the test set. Interestingly, the performance of the system did not change appreciably when using only entities tagged by the inbuilt tagger as compared to performance using the gold-tagged entities. The lack of an event anaphoric module, as well as problems in reducing events generated by a large trigger class to the task-specific event subset, were likely major contributory factors to the rather moderate performance.

1 Introduction

The Cancer Genetics (CG) task of the BioNLP-13 shared task (Pyysalo et al., 2013) has event types defined from a strict subset of GO biological processes. However, the events in the CG task have arguments that span a range of entities from molecules to system-wide processes, the latter focused primarily on cancer. Thus the CG task is an interesting case-study for text mining from a biological point of view, in that the task spans the literature from molecular events to behaviors linked to phenotypes, and thus considers a broader context than earlier BioNLP shared tasks (Kim et al., 2009, 2011).

An early article by Swanson (1988) explored the value of literature-based discovery (LBD) in discovering relations that span scientific sub-specializations. The LBD program of Swanson involves 3 nominally independent subtasks: (i) accurate representations of events within a document (b) normalization of entities to a standard representation to facilitate inter-document spanning and (c) a strategy to span event graphs across multiple documents. We explored the CG task primarily in the context of subtask (a) of this LBD program.

2 Methods

Our system currently consists of the following major components (a) Cocoa, a NER module that detects over 20 biomedical entity classes, including macromolecules, chemicals, protein/DNA parts, complexes, organisms, processes, anatomical parts, locations, physiological terms, parameters, values, experimental techniques, surgical procedures, and foods and (b) Peaberry, a 'stitcher' that combines local predicate-argument structures to produce a dependency-parse like output. The system also resolves sortal/pronominal anaphora and coreferences.

2.1 Entity detection

As entity detection is not part of the CG task, we provide only a brief overview of this module. However, as we did not use the entities provided by the event organizers on the test set, this description may be of interest given that our results with and without gold entities on the development and test sets are comparable (please see the Results section below).

The Cocoa entity detection system consists of the following modules run as a pipeline: (a) sentence boundary detection (b) acronym detection (c) a POS tagger based on Brill's tagger, postmodified for the biological domain (d) a fnTBL-based chunker, also heavily postmodified for the biomedical domain (e) an entity tagging module,

Proceedings of the BioNLP Shared Task 2013 Workshop, pages 86–93,
Sofia, Bulgaria, August 9 2013.

driven by dictionaries based both on words as well as morphological features, primarily prefixes and suffixes for biomedical entities, but also using infixes for chemical entities (f) entity tag based correction of chunks, primarily mistagged VP chunks (g) a narrow context/trigger based tagging of entities that are orthographically defined (presence of caps or numbers) such as assigning a protein tag for Cx43 from the phrase 'phosphorylation of Cx43' (h) a multi-word entity aggregator (i) a shallow coordination module for NPs (j) a limited set of hypernymic and appositional relations, followed by reuse of tags for orthographically defined unlabeled entities (k) a chemical formula detector. The entity tagger performs reasonably against proteins, anatomical parts and diseases as evaluated against existing tagged datasets (RelAgent, 2012).

2.2 Event Detection

The main steps here are: (a) detecting voice/finiteness of verbs (b) predicate-argument structure extraction for trigger words (c) argument merging and discourse connective parser (d) anaphora detection (e) discourse-connective based filling of empty themes and (f) sense disambiguation (WSD) of trigger words based on argument structure. A block-level pipeline of the system is given in Figure 1.

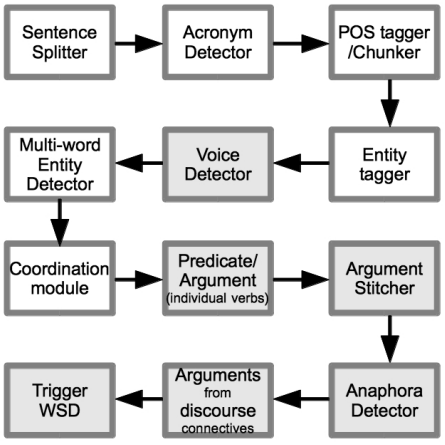

Figure 1. Block level pipeline of the system. Blocks with a light gray background are part of the event detection system (Peaberry), and are discussed here. The other blocks are part of the Cocoa entity tagger. WSD = Word sense disambiguation.

We will use a single sentence throughout to illustrate processing by the various modules:

"Concomitantly, immunostaining for apoptosis inducing factor (AIF) showed a time-dependent translocation from the mitochondria to the nucleus."

2.2.1 Voice detection

The voice detection module uses about 150 rules to detect the voice of a verb. It also classifies the verb as finite/nonfinite while marking its presence in a reduced or finite relative clause. The module determines these various aspects of a verb primarily with the local context, but uses the aspects of a previous verb in cases of coordinated verbs. Voice detection is facilitated by specific handling of (a) middle verbs, which appear to be in the active voice, but whose theme is the subject ('The protein translocated to the nucleus') (b) ergative verbs, which act like middle verbs when they do not have a direct object, but behave regularly when they are used transitively ('Protein levels increased' vs 'Application of the chemical increased protein levels') (c) intransitives, which are verbs that do not take a direct object, but whose subject is the agent ('The patient fell'), (d) verbs in the active voice, but with an object separated from the verb by a preposition ('leads to', 'resulted in', 'binds to'). Voice markup is therefore determined primarily by the roles of the subject/object, and is thus a little different from the voice markings as conventionally defined.

In the sample sentence, there is only one verb, and the output reads:

"[Concomitantly AV] , [immunostaining NP] for [apoptosis inducing factor (AIF) NP] [showed VP_Af] [a time -dependent translocation NP] from [the mitochondria NP] to [the nucleus NP] ."
where the verb phrase 'showed' is in the active voice ('A') and is finite ('f'). Another sentences better illustrates a wider range in voice markings:

[Adult naive T cells NP] , which [are VP_Pfcr] at [rest NP] in [normal conditions NP] , [proliferate strongly VP_Pf] when [transferred VP_Pnd] to [lymphopenic hosts NP] .

Here 'VP_Pfcr' stands for passive voice ('P'), finite ('f'), copula ('c'), and relative clause ('r'), while 'VP_Pnd' stands for Passive ('P'), non-finite ('n') and reduced ('d').

2.2.2 Argument extraction

Local arguments are extracted for all verbs in a sentence, as well as all nominals marked as potential triggers by the entity tagger. Currently, there are approximately 60 classes of predicate-argument structures based both on the particular prepositions heading noun phrases as well as entity tags; these classes cover about 500 specific trigger words. Additionally, there are generic argument structures for verbs and nominals not covered in the specific classes above. We accommodate 3 additional arguments apart from the agent/theme, such as FromLoc, ToLoc and AtLoc for movement-type trigger words. In addition, we also mark the subject/object nature of the arguments.

The argument structures for the sample sentence are shown in a pipe separated format (verb| cause|theme):

immunostaining | - | apoptosis inducing factor (AIF)
showed | immunostaining | a time -dependent translocation
-dependent | time | translocation
translocation | - | - | FromLoc:the nucleus| ToLoc: the mitochondria

2.2.3 Argument stitching and connectives

We link argument structures for individual triggers by looking for missing syntactical constituents for verbs (subject/object) or semantic constituents for nominals (agent/theme). For verbs, we use the voice/finite aspects of the current verb to locate previous verbs with which the current verb is associated with, either through embedding or by coordination. For example, in the sentence fragment: '... had no effect on the ability of beta-adrenergic agonists to stimulate internalization of beta2ARs , but blocked the ability of ...", 'blocked' coordinates with the finite verb 'had' but not with the non-finite 'to stimulate'. An example of an embedding is: 'With major interfering currents inhibited, NaCaEC was measured as the current that is sensitive to the nickel (Ni) during a descending voltage ramp.'. Here the VP 'was measured' is finite, and this allows its object 'the current' to be identified as the subject of 'is sensitive'.

Other examples of rules for resolving the arguments of relative clauses (RCs) are: (a) Discourse connectives ('whereas, 'whereby, 'because') form clausal boundaries and should not be crossed (b) Certain coordination markers ('besides', 'via') also should not be crossed for RC's (c) If an RC is recognized as coordinated with a prior RC, the arguments are transferred.

A general point in inferring missing arguments is that the nature of the current trigger word can also determine the nature of the induced argument. Certain trigger words ('induce', 'cause', enhance', 'prevent') can take an event as an argument , although most trigger words do not (theme argument for 'methylation'). Triggers in the former class are primarily regulatory actions and/or belief statements, which can take a clause or a nominal as an argument. The distinction between these two types of trigger words is related to that between 'embedding propositions' and 'atomic propositions' noted in Kilicoglu and Bergler (2012). An example is: 'Promoter methylation may interfere with AP1 binding to the promoter to cause aberrant Cx43 gene expression.', where it is the interference that causes aberrant expression.

The stitcher/parser does not examine the internal structure of chunks to locate missing arguments for predicates. This rule is violated for trigger words that can accept events as arguments, where the presence of an event trigger (as marked by the NER tagger) inside a NP is checked for. While this makes the process in some sense 'domain-neutral', it may also introduce errors unless the predicate-argument rules are complete and comprehensive for individual triggers.

The parser also locates discourse connectives ('whereas', 'because' , 'via', 'when') and assembles argument frames for these connectives, based on finiteness of verbs when possible. Connectives ('by') that can take nominals as arguments ('Localization ... by fusing') are also handled by the parser. Hypernymic and appositional relations are also detected at this stage. A final check locates all unattached prepositional phrases in the sentence and attaches them as verbal phrases to the nearest verb in a greedy step. At any point, the parser looks back no more than 2 verbs back for resolution, with parse time thus ~ O(2x), where 'x' is the number of trigger words in a sentence.

We recognize that the description of the 'stitching' process above is somewhat brief, but feel that a full description may not be appropriate here due to the large number of rules and interdependencies in the system. We note that: (a) the final output of the process is similar to a dependency parse, except that semantic roles are identified (b) the stitching is done in a shallow manner,

with two verbs look-back at most, and is hence reasonably fast and (c) the implementation is our own, and does not borrow from existing parsers. We plan to describe this system in greater detail in a separate publication elsewhere.

As an example, in the paraphrase 'X activates Y to increase Z', the arguments are:

activates | X | Y
increase | - | Z

and the stitcher recognizes the infinitival 'to' construct, and transfers the previous event as the agent for 'increase':

activates | X | Y
increase | activates | Z

2.3 Anaphora

We implemented the algorithm of Lappin and Leass (1994) for pronominal anaphora, as implemented by Kennedy and Boguraev (1996), with additional weights for matching entity tags for headwords. The weights were refined against handpicked abstracts, but are yet to be completely validated. In addition, we also resolved sortal anaphora ('this protein', 'these genes') and prenominal anaphora ('its binding partner', 'their properties') by the same rules as used for pronominal anaphors ('it', 'they'), but with different weights. We also implemented event anaphora, i.e. reference of one trigger word to another trigger word with the same root (lemma) or another event in the same class (for regulation triggers). Due to lack of time, we could not completely test the performance of event anaphora, and they were dropped in the test set. Coreference resolution with the determiner 'the' ('the gene') was not implemented.

2.4 Transferring arguments across events

Certain arguments can be resolved by comparing argument structures for events linked by discourse connectives (DCs), such as :

'found to overexpress eph mRNAs without gene amplification' (DC: 'without')

'Upon retroviral transduction of the mouse c-myc gene, Rat 6 cells showed mildly altered morphology' (DC: 'Upon')

'SCAI acts on the RhoA-Dia1 signal transduction pathway and localizes in the nucleus, where it binds and inhibits the myocardin-related transcription factor MAL by forming a ternary complex with serum response factor (SRF).' (anaphoric resolution for 'it' followed by a discourse connective:'by')

When events are linked by a discourse connective, arguments can be transferred if the events are in the same event class. Even if the events are of different classes, the theme can be transferred if it satisfies the entity type constraints of the recipient event. Further, certain belief/demonstration trigger words ('display', 'show', 'exhibit', 'demonstrate') that take an event as the theme have a similar structure: 'Cloning of a human phosphoinositide 3-kinase with a C2 domain that displays reduced sensitivity to the inhibitor wortmannin.' or 'X exhibits cytotoxicity against cell lines'. Agent arguments for such verbs are transferred to the appropriate argument slot of the theme event. In certain contexts, verbs such as 'act' which can take an infinitival 'to' complement behave similarly: 'p15 may act as an effector of TGF-beta-mediated cell cycle arrest.'.

For the sample sentence, the trigger/belief word 'showed' causes a transfer of the theme slot of its cause process ('immunostaining', a Planned_process in the CG task) to the same slot in the theme event ('translocation'):

immunostaining | - | apoptosis inducing factor (AIF)
showed | immunostaining | a time -dependent translocation
-dependent | time | translocation
translocation | - | apoptosis inducing factor (AIF) | FromLoc:the nucleus| ToLoc: the mitochondria

2.5 Runtimes

The run-time of the system is about 100 ms/sentence on a 2007 vintage dual-core system. This time was estimated by processing whole abstracts varying from 10-15 sentences. This figure includes the time for all components, including entity recognition, parsing, intra-document anaphora resolution (both sortal/pronominal and event), event extraction and final A1/A2 output. The extrapolated time of processing for the entire Medline corpus (1.2 x 10^8 sentences in 2013) is about 180 CPU-days.

3 Results

We first tested the system against the development set by using the internal entity detector (Cocoa) to tag entities, and using these tags alone till the end of the event extraction phase, and only then remapping the Cocoa-tagged entities to

entities in the gold annotations ('a1' entities) given by the task organizers. This gave a score (f-measure) of 52.2% with the evaluation options '-s -p' which stand respectively for soft span matching and partial recursive matching. We then reran the event extraction module after removing all internally generated entity tags for chemicals, proteins and anatomical parts and tagging only such entities as were specified in the gold 'a1' files. To our surprise, the f-measure was 2% lower on the development set when using the gold entities. This probably indicates an unwanted dependence of the event extraction module on some peculiarities in the way the internal Cocoa tagger tags entities. We are currently analyzing the results for such dependencies (see Discussion for some examples). Nevertheless, the results are encouraging in that the system performance is similar with or without reference entities and thus may be indicative of performance on a new document collection where entities are not specified manually beforehand.

As the task allows only one submission, we submitted the results of the system with entities tagged by the internal tagger and mapped only at the end to the gold tagged entities. This was based on the better performance of this approach against the development set. However, the results of the system were considerably lower on the test set (f = 45.3%; best score by TEES 2.1 system = 55.4%; Pyysalo et al., 2013). Using the evaluation portal for the test dataset, the results with gold-tagged entities improved the performance only by 0.3%, confirming that, at least at the performance levels of this system, the inbuilt Cocoa entity tags can substitute for pre-annotated entities.

The performance on the test set was low primarily against the events in the regulation class (f=35.6%), which form about 40% of the events in both the test and development sets. This is similar to the result in the development set, where the performance in the regulation class was also quite low at f=37%. Part of the reason for this is that the system's rules for regulatory triggers generally give preference to other events over entities as causes/agents. Thus for example in the sentence fragment (PMID:21963494) 'AglRhz induced activation of caspase-3 and poly(ADP-ribose) polymerase (PARP), and DNA fragmentation in HT-29 cells, leads to induction of apoptosis as well as suppression of tumorigenicity of HT-29 cells.', the gold annotations state that 'AglRhz' is the cause for the trigger word 'leads', while the Peaberry system prefers the trigger word 'induced' for the causative agent. However, we have not done a detailed study to examine if such differences account for more than a small minority of the errors that contribute to low performance in the regulatory class. Overall, and surprisingly for a rule-based system, the precision was quite low on both the test set (49%) and the development set (54%). The low overall precision was dominated by the corresponding number for regulatory events (37% and 44% on test and development sets respectively), but the precision of non-regulatory events was quite dismal as well (please see Discussion section below).

The low recall for regulatory events can be caused by low recall for those primary (i.e. non-regulatory) events that are regulated. In the development set, the recall for these non-regulatory classes varied between 55% and 75%, but in the test set the recall for some primary event classes (Pathology and General event classes) dropped to ~30-40% (see Table 1 below). Another reason for low recall is the absence of themes for primary events when these themes are lifted/transferred from mentions of the same trigger word in previous sentences. Our lack of a event anaphora module would thus certainly have contributed to the low recall for such primary events. We are analyzing the gold annotations to determine other causes for the low precision and recall in the development dataset.

Event Class	Recall	Precision	Fscore
Anatomy	63.34	80.29	70.82
Pathology	43.30	54.20	48.14
Molecule	57.46	64.38	60.72
General	34.67	49.82	40.89
Regulation	34.22	37.05	35.58
Modifier	26.24	37.50	30.88
Total	41.73	49.58	45.32

Table 1. Summary of results for the Test set. Recall, precision and F-score are shown for event classes for anatomical changes, pathology, molecular processing events, general events (binding and movement), regulatory events, modifiers (negation and speculation) and the total score.

4 Discussion

We have developed a rule-based linguistically motivated system for tagging entities and extracting events from biomedical documents. A major

problem with our linguistically-based system is the large open-ended number of trigger words that generate events. This explosive event generation occurs as the system generates predicate argument structures for all verbs in a document as well as for generically defined nominal processes (which are marked as event triggers by morphological considerations, such as words ending in "ation"). Moreover, the entity tagger also marks a variety of other words as event triggers when they are known to stand for biological or disease processes, in the Gene Ontology for example. Projecting the system output into a limited sets of trigger words for a particular task was somewhat problematic for us, although a good training exercise on transferring arguments (e.g. the theme) from 'other' trigger words into the subset of trigger words sufficient for the task. It is possible that defects in this argument transfer process could account for some of the low performance in the test set.

Developing a rule-based system involves a large amount of manual work in tuning the various aspects of the system to the task at hand. This is true even if the framework for the system is already in place. For example, with the CG task, the predicate-argument structures for each individual trigger have to be exhaustively worked out to handle all possible locations of argument structures. For certain triggers, the theme in the CG task is somewhat indirect, as in the sentence: 'Almost all patients respond to G-CSF with increased neutrophils, reduced infections, and improved survival.', where the theme of 'response' are not the patients but the 'increased neutrophils'. This is perhaps clearer in the paraphrase: "Organism responded to Drug with Symptoms", and cellular symptoms are the appropriate theme for the trigger 'responded' in the CG task. Distinguishing such a sentence from a syntactically similar but semantically distinct sentence ' Organism responded well to Drug' is a challenging, and perhaps arduous, task for a rule-based linguistically motivated system. We note that the CG task annotations are quite consistent in this aspect, as the theme is again Symptoms in the paraphrase 'Drug protects Organism from Symptoms'.

Further, in certain sentences, it is somewhat hard to express the meaning in the A2 notation. This is particularly true for adjectives which refer to the state of an entity rather than an event. Consider (PMID 17367752): "These results suggest that SWAP-70 may be required for oncogenic transformation and contributes to cell growth in MEFs transformed by v-Src." where one of the gold annotations transcribes functionally as

'contributes (Agent: SWAP-70, Theme: transformed (Theme: MEFs))'

which suggests that SWAP-70 contributes to the transformation of MEF's, whereas 'transformed' is only an attribute of the MEF's for this annotation. These aspects of the CG task annotations are particularly hard to capture in a rule-based system. A similar problematic sentence is 'recombinant EBVs that lack the BHRF1 miRNA cluster display a reduced ability to transform B lymphocytes in vitro' where the gold annotations read:

reduced (Agent: recombinant_EBVs Theme: transform (B_lymphocytes))

The sentence however suggests that it is the 'lack' of a 'miRNA cluster' in the EBV's that reduces the transformation. Again, this reading is somewhat hard to express in A2 notation.

As an additional example of the task complexity, we noted that distinguishing between the role of the trigger word 'transform' as 'Cell_transformation' and its role as a 'Planned_process' seems to require some level of discourse analysis at least in the CG training data.

Some defects in the system output arise from differences in interpretation. In the sentence 'Merlin protein might contribute to the initiation of metastasis of NSCLC.', (PMID:2174350), the gold annotations read:

'contribute(Agent:Merlin, Theme: initiation (Theme: NSCLC))'
'contribute (Agent: Merlin Theme: metastasis (Theme: NSCLC))

where NSCLC is a cancer. Peaberry gives instead

'contribute (Agent: Merlin, Theme: initiation(Theme: metastasis(Theme: NSCLC)))

As 'initiation' generally requires an event/process/disease as a theme, its theme could be either 'metastasis' or 'NSCLC', and the system makes a greedy choice in this case. As changes in this logic would have a system-wide impact, this example perhaps shows the inflexibility of the system.

A straightforward example shows the costs of missed anaphora: 'Gene silencing and over-ex-

pression techniques were used to modulate RASSF1C expression in human breast cancer cells.' The system misses both events 'expression (Theme:RASSF1C)' and 'over-expression (Theme: RASSF1C)', both themes resolving to the anaphoric entity 'Gene', which needs resolution. Similar considerations apply for the sentence: 'knockdown of HDGF, an up-regulated protein and a target of NF-kappaB, induced cell apoptosis', where 'protein' and not 'HDGF' is seen as the theme of the trigger 'up-regulated'.

Rule based systems have been used in previous BioNLP shared tasks. Such a system, described by Kilicoglu and Bergler (2012), was employed for the BioNLP shared task 2011. This system used output from the Stanford dependency parser together with the notion of embedding to construct a semantic graph, from which propositions were extracted. These propositions were converted into events, and semantic roles were derived depending on the nature of the predicate trigger word. In comparing the performance of this system on the 2011 GENIA task against our system on the CG task in common categories, the striking difference is that our precision is far lower in most categories (see Table 2), even while recall is comparable. In particular, the difference in precision in non-regulation categories is quite noticeable. We are yet to understand the reasons for these low precision scores in the Peaberry system.

Event Class	GENIA	CG
Localization	90.36	59.43
Binding	49.66	34.69
Gene expression	86.84	71.46
Transcription	58.95	100.00
Phosphorylation	94.56	70.83
Regulation	45.85	37.05
Modifier	40.89	37.50
Total	59.58	49.58

Table 2. Comparison of precision between two rule-based systems for similar event classes: (a) system of Kilicoglu and Bergler (2012) in the GENIA task of BioNLP 11 (b) current system in the CG task of BioNLP 13.

We noted in the Results section that performance of the system with and without gold-tagged entities (tagged in the latter case by the internal Cocoa tagger) was similar, 0.7% better with the gold entities in the test run, and 2% better with internal entities on the development set. A preliminary analysis shows that the reduction in some cases with gold entities was due to peculiarities in the way the system handles acronyms. The internal tagger lumps together an acronym with its expansion as a single token, while the gold annotations tokenizes the acronym and the definition separately. This affects downstream processing, especially in the stitching module. The gold annotations also do not markup sortal anaphors ('gene' in 'this gene'), and the system depends on entities being marked up in such anaphors to find a referent. Altogether, while the results may initially seem surprising, they do not support any notion that automatically predicted entities are somehow better than gold annotated entities for event extraction systems. At most, the similarities in results with and without gold annotated entities are indicative of a comparable performance, a very moderate f =~ 0.45, of the complete system on a new document collection without gold annotations.

We note that it seems possible that the rules developed for the CG task can be extended without major modifications to the PC and the GE tasks, whose set of event triggers are a subset of the CG task, without degrading the performance of the CG task. This may be one of the few advantages of a labor-intensive rule-based system; however, we are yet to validate such a supposition.

Cancer is founded at the molecular/genetic/cellular level and is localized to an individual organ/tissue before metastasis. It would thus seem that the text processing logic used for the CG task should be generalizable (at least) to diseases of individual organs. However, cancer is not a true multi-organ systemic problem of the type that characterizes life-style diseases such as diabetes and cardiovascular disease, which are both linked to multiple genomic loci as well as to multiple organs, and it would be interesting to explore coverage of event extraction schemes for these diseases with the text mining techniques developed in the CG task. In this context, we note that automatic annotation of all events in a document needs to be followed by highlighting of the novel events/properties in the document, which may require some discourse analysis.

Reference

C. Kennedy and B. Boguraev (1996) Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. COLING '96 Proceedings of the

16th conference on Computational linguistics. 1:113-118

H. Kilicoglu and S. Bergler 2012. Biological Event Composition. BMC Bioinformatics 13:Supplement 11. Edited by J-D. Kim, S. Pyysalo, C. Nedellec, S. Ananiadou and J. Tsujii.

J. D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In Proceedings of the Workshop on BioNLP: Shared Task. 2009:1-9.

J. D. Kim , S. Pyysalo, T. Ohta, R. Bossy, and J. Tsujii. 2011. Overview of BioNLP Shared Task 2011. In Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task. 2011:1-6.

S. Lappin and H. J. Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. J. Comp. Ling. 20:(4):535-561

S. Pyysalo, T. Ohta, M. Miwa, H.C. Cho, J. Tsujii J, and S. Ananiadou. 2012. Event extraction across multiple levels of biological organization. Bioinformatics. 28(18):i575-i581

S. Pyysalo, T. Ohta and S. Ananiadou. Overview of the Cancer Genetics (CG) task of BioNLP Shared Task 2013. In Proceedings of BioNLP Shared Task 2013 Workshop. To appear.

S. Pyysalo, T. Ohta and S. Ananiadou. 2013. Cancer Genetics task. Final evaluation results - RelAgent. http://weaver.nlplab.org/~bionlp-st/BioNLP-ST-2013/CG/final-results/RelAgent.html

RelAgent. 2012. Evaluation of Cocoa against some corpora. http://npjoint.com/CocoaEval.html

D. Swanson. 1988. Migraine and Magnesium: Eleven Neglected Connections. Persp. Bio. Med. 31(4):526–557.