The StringNet Lexico-Grammatical Knowledgebase and its Applications

David Wible **Nai-Lung Tsao**
National Central University
No.300, Jhongda Rd.
Jhongli City, Taoyuan County 32001, Taiwan
wible@stringnet.org beaktsao@stringnet.org

Abstract

This demo introduces a suite of web-based English lexical knowledge resources, called StringNet and StringNet Navigator (http://nav.stringnet.org), designed to provide access to the immense territory of multiword expressions that falls between what the lexical entries encode in lexicons on the one hand and what productive grammar rules cover on the other. StringNet's content consists of 1.6 billion hybrid n-grams, strings in which word forms and parts of speech grams can co-occur. Subordinate and super-ordinate relations among hybrid n-grams are indexed, making StringNet a navigable web rather than a list. Applications include error detection and correction tools and web browser-based tools that detect patterns in the webpages that a user browses.

1 Introduction and Background

This demo introduces a suite of web-based English lexical knowledge resources, called StringNet and StringNet Navigator (http://nav.stringnet.org), which have been designed to give lexicographers, translators, language teachers and language learners direct access to the immense territory of multiword expressions, more specifically to the lexical patterning that falls in the gap between dictionaries and grammar books.

MWEs are widely recognized in two different research communities as posing persistent problems, specifically in the fields of computational linguistics and human language learning and pedagogy.

In computational linguistics, MWEs are notorious as a "pain in the neck" (Sag et al 2002; Baldwin et al 2004; Villavicencio et al 2005; inter alia). The high proportion of MWEs with non-canonical structures lead to parse failures and their non-compositional or only partially compositional semantics raise difficult choices between which ones to store whole and which ones to construct as needed. Perhaps above all, this massive family of expressions resists any unified treatment since they constitute a heterogeneous mix of regularity and idiomicity (Fillmore et al 1988).

The other area where they famously cause difficulties is in human language learning and teaching, and largely for reasons parallel to those that make them hard for NLP. They resist understanding or production by general rules or composition, and they constitute an unpredictable mix of productivity and idiomicity.

The StringNet lexico-grammatical knowledge-base has been designed to capture this heterogeneity of MWEs by virtue of its unique content and structure. These we describe in turn below.

2 StringNet Content: Hybrid N-grams

The content of StringNet consists of a special breed of n-grams which we call hybrid n-grams (Tsao and Wible 2009; Wible and Tsao 2010). Unlike traditional n-grams, there are four different categories of gram type. From specific to general (or abstract) these four are: specific word forms (*enjoyed* and *enjoys* would be two distinct word forms); lexemes (**enjoy**, including all its inflectional variations, *enjoyed*, *enjoys*, etc); rough POS categories (V, N, etc); and fine-grained POS categories (verbs are distinguished as VVn, VVd, VVt, etc.). A hybrid n-gram can consist of any sequence from any of these four categories with

Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 128–130,
Portland, Oregon, USA, 23 June 2011. ©2011 Association for Computational Linguistics

our stipulation that one of the grams must be a word form or lexeme (to insure that all hybrid n-grams are lexically anchored). A traditional bi-gram such as *enjoyed hiking* can be described by 16 distinct hybrid n-grams, such as *enjoyed VVg*, ***enjoy** VVg*, ***enjoy hike***, and so on. A traditional 5-gram, such as *kept a close eye on* has 1024 hybrid n-gram variants (4^5), e.g., *keep a close eye on*; *kept a [Adj] eye on*; *keep a close [N][Prep]*; and so on. We have extracted all hybrid n-grams ranging in length from bigrams to 8-grams that are attested at least five times in BNC. StringNet's content thus consists of 1.6 billion hybrid n-grams (including traditional n-grams), each indexed to its attested instances in BNC.

3 Structure and Navigation

Rather than a list of hybrid n-grams, StringNet is a structured net. Hybrid n-grams can stand in sub-ordinate or super-ordinate relation to each other (we refer to these as parent/child relations). For example, the hybrid tri-gram *consider yourselves lucky* has among its many parents the more inclusive *consider [prn rflx] lucky*; which in turn has among its parents the even more general *consider [prn rflx] [Adj*] and *[V] [prn rflx] lucky* and so on. We index all of these relations within the entire set of hybrid n-grams.

StringNet Navigator is the Web interface (shown in Figure 1) for navigating this massive, structured lexico-grammatical knowledgebase of English MWEs. Queries are as simple as submitting a Google query. A query of the noun *trouble* immediately shows users (say, language learners) subtle but important patterns such as *take the trouble [to-V]* and *go to the trouble of [VVg]* (shown in Figure 2). Submitting *mistake* yields *make the mistake of [VVg]* and *it would be a mistake [to-V]*. StringNet Navigator also accepts multiword queries, returning all hybrid n-grams where the submitted words or the submitted words and POSs co-occur. For all queries, clicking on any pattern given in the results will display all the attested example sentences with that pattern from BNC. Each listed pattern for a query also gives links to that pattern's parents and children or to its expansion (longer version) or contraction (shorter version) (See Figure 2).

4 Some Applications

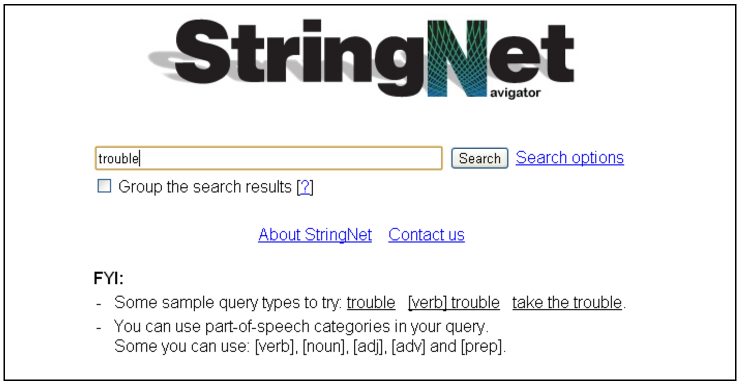
Among the many sorts of knowledge that StringNet renders tractable is the degree of frozenness or substitutability available for any MWE. Thus, not only does a query of the noun *eye* yield the string *keep a close eye on.* Navigating upward reveals that *close* and *eye* in this string can be replaced (*keep a close watch on*; *keep a careful eye on*; *keep a tight grip on*; *keep a firm hold on*, etc), but also that, in this same frame *keep a [Adj][N] on*, the verb slot occupied by *keep* is basically unsubstitutable, essentially serving as a lexical anchor to this expression. Thus, due to its structure as a net, StringNet makes it possible to glean the degree and location(s) of the frozenness or substitutability of an MWE.

4.1 Error Checking

Automatic error detection and correction is a rapidly growing area of application in computational linguistics (See Leacock et al 2010 for a recent book-length review). StringNet supports a novel approach to this area of work. The flexibility afforded by hybrid n-grams makes it possible to capture patterns that involve subtle combinations of lexical specificity or generality for different grams within the same string. For example, running StringNet on BNC data shows that 'enjoy hiking' is best captured as an instance of the lexeme **enjoy** followed by a verb in –ing form: ***enjoy** Vvg*. For error checking this makes it possible to overcome sparseness. Thus, while BNC has no tokens of either 'enjoy spelunking' or 'enjoy to spelunk,' we can distinguish between them nevertheless and detect that the former is correct and the latter is an error. The wide range of error types that can be handled by a single algorithm run on StringNet will be shown in the demo.

4.2 Browser-based Tools

Other tools include a toolbar that can be installed on the user's own web browser (Wible et al 2011), from which the system can detect lexical patterns in the text of the web pages the user freely browses. A "Query Doctor" on the toolbar detects errors in multiword queries (submitting 'in my point of view' triggers the suggestion: 'from my point of view').



trouble | Search | Search options

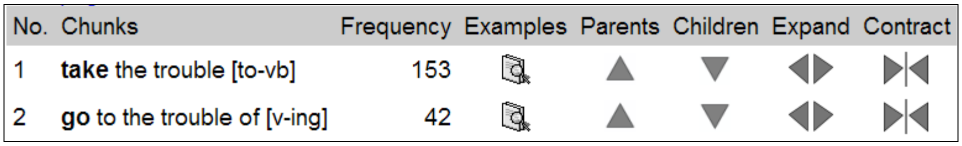
□ Group the search results [?]

About StringNet Contact us

FYI:

- Some sample query types to try: trouble [verb] trouble take the trouble.
- You can use part-of-speech categories in your query.
  Some you can use: [verb], [noun], [adj], [adv] and [prep].

Figure 1: StringNet Navigator front page.



No.	Chunks	Frequency	Examples	Parents	Children	Expand	Contract
1	**take** the trouble [to-vb]	153	🔍	▲	▼	◀▶	▶◀
2	**go** to the trouble of [v-ing]	42	🔍	▲	▼	◀▶	▶◀

Figure 2: Top 2 search results for "trouble"

5 Conclusion

Future areas of application for StringNet include machine translation (e.g., detecting semi-compositional constructions); detection of similar and confusable words for learners, document similarity using hybrid n-grams as features, and StringNet Builder for generating StringNets from corpora of languages other than English and from domain-specific corpora.

Acknowledgments

The work described in this paper was partially supported by the grants from the National Science Council, Taiwan (Project Nos. 99-2631-H-008-004- and 98-2511-S-008-002-MY2).

References

Timothy Baldwin, Emily M. Bender, Dan Flickinger, Ara Kim and Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, pp. 2047-2050.

Charles J. Fillmore, Paul Kay, and Mary Katherine O'Connor. 1988. Regularity and Idiomaticity in Grammatical Constructions: the Case of *Let Alone. Language* 64: 501–538.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault, 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool Publishers.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Mexico City, Mexico, pp. 1-15.

Nai-Lung Tsao and David Wible. 2009. A Method for Unsupervised Broad-Coverage Lexical Error Detection and Correction. *The NAACL Workshop on Innovative Use of NLP for Building Educational Applications*, Boulder, Colorado, June 2009.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the Special Issue on Multiword Expressions: Having a Crack at a Hard Nut. *Computer Speech & Language* 19(4): 365-377.

David Wible and Nai-Lung Tsao. 2010. StringNet as a Computational Resource for Discovering and Investigating Linguistic Constructions. *The NAACL Workshop on Extracting and Using Constructions in Computational Linguistics*, Los Angeles, June 2010.

David Wible, Anne Li-E Liu and Nai-Lung Tsao. 2011. A Browser-based Approach to Incidental Individualization of Vocabulary Learning. *Journal of Computer Assisted Learning*, in press, early view.