SimConcept: A Hybrid Approach for Simplifying Composite Named Entities in Biomedical Text

Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu

***Abstract*—One particular challenge in biomedical named entity recognition (NER) and normalization is the identification and resolution of composite named entities, where a single span refers to more than one concept (e.g., BRCA1/2). Previous NER and normalization studies have either ignored composite mentions, used simple *ad hoc* rules, or only handled coordination ellipsis, making a robust approach for handling multitype composite mentions greatly needed. To this end, we propose a hybrid method integrating a machine-learning model with a pattern identification strategy to identify the individual components of each composite mention. Our method, which we have named SimConcept, is the first to systematically handle many types of composite mentions. The technique achieves high performance in identifying and resolving composite mentions for three key biological entities: genes (90.42% in F-measure), diseases (86.47% in F-measure), and chemicals (86.05% in F-measure). Furthermore, our results show that using our SimConcept method can subsequently improve the performance of gene and disease concept recognition and normalization. SimConcept is available for download at: http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/SimConcept/**

***Index Terms*—bioNLP, coordination ellipsis, composite mentions, named entity recognition, text mining.**

I. Introduction

In biomedical text mining, many studies have focused on automatically extracting relevant information from published literature. The relevant information is commonly focused on a specific topic, such as protein–protein interactions [2], [3], protein transport and localization [4]–[6], drug-disease associations [7]–[9], or gene function extraction [10]. Most of the common retrieval methods apply natural language processing or machine learning to identify relations in text. One crucial step toward this goal is automatically recognizing bioconcept mentions (e.g., gene/protein)—the task of named entity recognition (NER)—and mapping the bioconcept to a specific database identifier (e.g., NCBI EntrezGene)—the task of normalization. Many international biomedical text mining competitions (e.g., BioCreative) have, therefore, focused on these tasks [11]–[13]. Genes, diseases, and chemicals are particularly notable for not only being important concepts, but also being the most popular concepts in biomedical literature search [14], [15]. Most normalization studies face two challenges: term variation and ambiguity [16]–[22]. Many previous studies have defined individual strategies (e.g., machine learning, statistical inference, and rule-based methods) to deal with these two issues. However, a particular type of error that has not been handled well is composite mentions, where a single span refers to more than one concept (e.g., "SMADs 1, 5, and 8"). Such mentions specifically refer to multiple concepts; they are, thus, distinct from phenomena such as protein complexes and chemical mixtures where multiple entities combine to form a single physical unit. We observe that in our datasets, approximately 10% of gene, disease, and chemical mentions are composite mentions, hence, it is important to handle them properly. This study presents a new method for bioconcept mention simplification in a systematic fashion.

Most related previous studies have focused on text simplification (including both document/paragraph [23]–[27] and sentence [28]–[31] levels). The few studies that have considered mention simplification have only addressed coordination ellipsis. Buyko *et al.*, [32] developed a CRF-based method with three states: conjunction, conjuncts, and ellipsis antecedent. For example, in "human and mouse cells," "human" and "mouse" are conjuncts, "and" is a conjunction, and "cells" is an ellipsis antecedent. They evaluated their method using the GENIA [33] corpus, obtaining 86% accuracy. Due to the lower performance of this method on complex ellipsis (e.g., "recombinant human nm23-H1, -H2, mouse nm23-M1, and -M2"), Chae *et al.*, [34] developed a pattern-based method, using lexicons to identify the region of each component (i.e., conjunction, conjuncts, and ellipsis antecedent) for each mention. However, these previous studies have focused on only one type of composite mention: mentions with coordination ellipsis.

In this study, a total of six types of composite mentions are considered including five distinct types (including abbreviation pair) and a mixed type of mentions.

1) *Mention with coordination ellipsis:* the concepts in this type of mentions share part of the mention region, such as the token "SMAD" in "SMADs 1, 5, and 8."
2) *Range mention:* Like mentions with coordination ellipsis, these mentions share part of the mention region, however, this type represents a range of entities rather than a discrete set (e.g., "SMAD 2 to 4").
3) *Individual mention:* this is an independent composite mention. All concepts can be separated into nonoverlapping spans (e.g., "BTK/ITK/TEC/TXK").
4) *Overlap abbreviation pair mention:* The long form and short form share some tokens, like "COUP (chicken ovalbumin upstream promoter) transcription factor" where the phrase "transcription factor" is shared across "COUP tran-

Manuscript received December 15, 2014; revised April 2, 2015; accepted April 5, 2015. Date of publication April 13, 2015; date of current version July 23, 2015. This work was supported by the NIH Intramural Research Program, National Library of Medicine. This paper [1] was presented at the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, September 2014.

The authors are with the National Institutes of Health and National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD 20894 USA (e-mail: chih-hsuan.wei@nih.gov; robert.leaman@nih.gov; zhiyong.lu@nih.gov).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2015.2422651

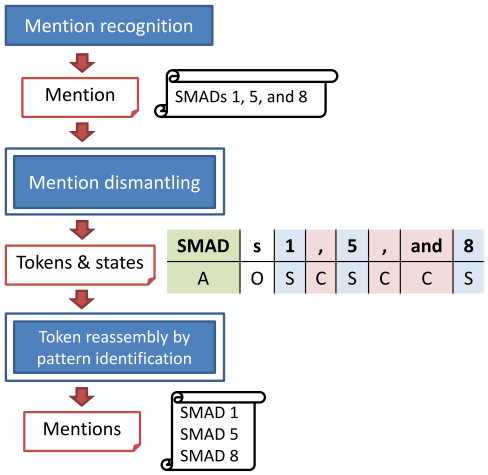


Fig. 1. Architecture of the SimConcept mention simplification method.

scription factor" and "chicken ovalbumin upstream promoter transcription factor." But the two concepts indicate the same database identifier.

5) *Individual abbreviation pair mention:* this is an independent composite mention where the two individual concepts indicate the same database identifier (e.g., "ectodermal dysplasia").
6) *Mixed mention:* It is a mixed mention of any two above types, like "high mobility group protein 1 and 2"—a mix of type 1 and 4.

The three main contributions of this study are: 1) a new tool called SimConcept was developed to handle six types of composite mentions, more than any other methods previously reported; 2) when applied to the three bioconcepts (i.e., gene, disease, and chemical), our method achieved state-of-the-art performance; and 3) based on our success on more than one entity type, our approach is shown to be robust and generalizable.

II. Methods

Overall, our method consists of two modules as shown in Fig. 1. The first module consists of a conditional random field (CRF) model. In this module, the input mention is separated into tokens and each token assigned labels according to the most likely sequence of states through the model. The second module reassembles the tokens into individual mentions using a pattern identification method.

A. CRF Model

As mentioned earlier, we regarded this mention simplification problem as a sequence-labeling task. To recognize the composite mentions, we observed the composition of those mentions and defined nine states for building a CRF model [35]: antecedent (A); strain/suffix(S); conjunction of mentions with coordination ellipsis (C); conjunction of range mentions (C_R); left parentheses of abbreviation pair (L); right parentheses of abbreviation (R); right parentheses of abbreviation, but the abbreviation and long form cannot be separated (R_o); conjunction of individual mentions (I); Redundant (O). The states "C," "$C_{R,}$" "L," "R," "R_o" (L and R/R_o occur in pairs), and "I" are conjunction states that can use to recognize the mention types. If one mention includes two or more conjunction states, this mention would be identified as a mixed mention.

Our implementation uses a linear chain CRF [35] provided by CRF++ (http://crfpp.googlecode.com/svn/trunk/doc/index.html). CRF++ applies L-BFGS [36], which is a quasi-Newton algorithm for large scale numeric optimization problems.

B. CRF Features

We used tmVar's tokenization [37] and part of its features in SimConcept development. Like tmVar, our tokenization separates uppercase characters, lowercase characters, and digits. For example, "SMADs 2 to 4" is separated to "SMAD," "s," "2," "to" and "4." We adapted tmVar's features to reflect the difference in input between tmVar (i.e., documents) and SimConcept (i.e., individual mentions). After reviewing the evidence for different token types of a mention, we defined several suffixes, prefixes, and some semantic types for identifying bioconcepts (i.e., gene, disease, and chemical) mention characteristics. In particular, most mention suffixes for disease and chemical mentions are not digits, for example, "breast and ovarian cancer" (disease) and "b-sitosteryl and stigmasteryl linoleates" (chemical), which might be difficult to recognize without semantic evidence. Therefore, we collected the semantic features used in some previous studies [37]–[39] and grouped the suffixes/prefixes we defined into semantic feature types such as those shown as follows.

1) *Chemical Suffix:* yl, ylidyne, oyl, sulfony, one, and etc.
2) *Chemical Alkane Stem:* meth, eth, prop, tetracos.
3) *Chemical Trivial Ring:* benzene, pyridine, toluene.
4) *Chemical Simple Multiplier:* di, tri, tetra, and etc.
5) *Chemical elements:* hydrogen, helium, lithium, carbon, and etc.
6) *Disease Suffix:* cancer, disease, symptom, and etc.
7) *Gene/Protein Suffix:* gene, protein, receptor, unit and etc.
8) *Family, Complex :* family, subfamily, superfamily, complex.

We also continue to use three of tmVar's features types (i.e., character features, case pattern features, and contextual features). Character features include number of digits, number of uppercase and lowercase letters, number of all characters, and specific characters (;,. – > + _ / ?). Case pattern features are created by replacing uppercase alphabetic character to "A" and any lower case to "a." Likewise any number (0–9) is replaced by "0." Moreover, we also merged consecutive letters and numbers to generate additional features, such as "AAA" to "A." Next, we used evidence in full text as a feature. That is, we would search candidate mentions in full text and look for their presence. For example, in determining how to decompose "W1 and W2 W3," we would search the bigram "W1 W3" in full text. If found, it is reasonable to infer that "W1 W3" is a valid and meaningful mention (e.g., ovarian and breast cancer). Otherwise, it is more likely that W1 should be separated by itself (e.g., emphysema and liver disease). Finally, in order to take advantage of contextual

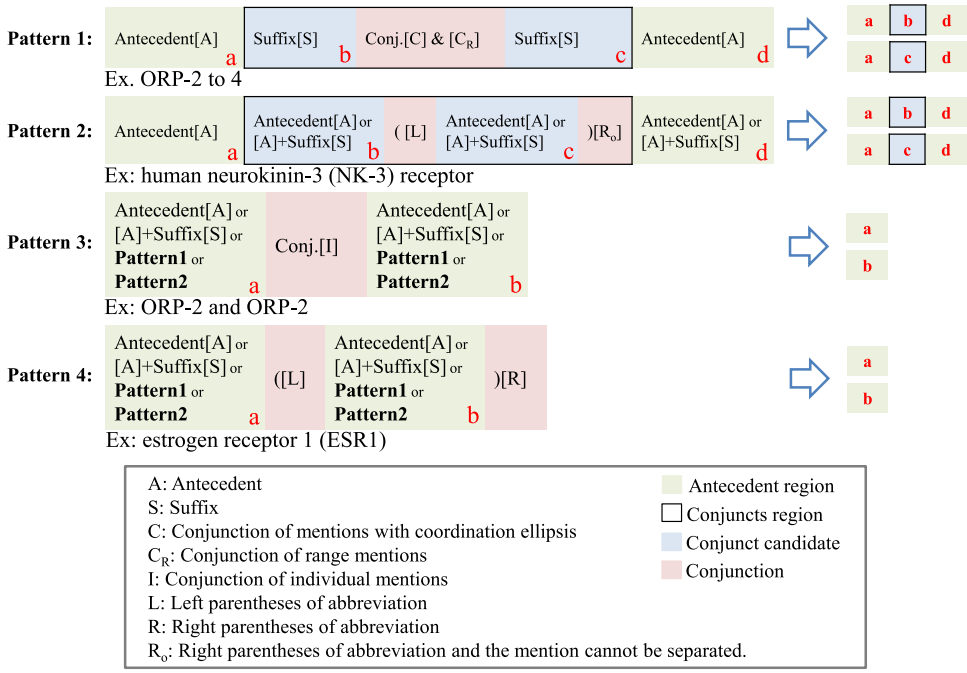


Fig. 2. Patterns for formulating bioconcept mentions. Note that in Patterns 1 & 2, it is common to have Antecedent appearing either at the beginning or the end but not both locations (i.e., component *a* and *d* may not appear together in one mention).

information, for a given token we included the token and semantic features of three neighboring tokens from each side.

C. Token Reassembly Through Pattern Identification

By observing the characteristics of composite mentions in our training data, we manually defined four patterns to model the six types of composite bioconcept mentions, as shown in Fig. 2. To simplify mentions, we distinguish between the antecedent region (green), conjuncts region (frame), conjunct candidate (blue), and conjunctions (red). The tokens in antecedent region should be present in all possible mentions. The tokens in conjuncts region should be replaced by all possible conjunct candidates in this region. Every conjuncts region consists of at least one conjunction. Conjunctions are used to separate individual conjunct candidates. In our definition, every mention can map to one of the patterns. Range mentions and mentions with coordination ellipsis map to Pattern 1. As shown in Fig. 3(a), the "ORP-2 to -4" is a range mention that can be separated to "ORP-" (antecedent region), and "2 to 4" (conjuncts region). In conjuncts region, all possible candidates (i.e., 2, 3, and 4 in "2 to 4") belong to one of the possible mentions. Therefore, "ORP-2 to -4" is reassembled to "ORP-2," "ORP-3," and "ORP-4." In another similar case, the "ORP-1 and -2" is similar to "ORP-2 to -4". The major difference is the conjunction (i.e., "and"). In this case, "-1" and "-2" in conjuncts region are independent. Therefore "ORP-1 and -2" becomes "ORP-1" and "ORP-2."

Observing these two cases, it becomes clear that the difference between range mentions and mentions with coordination ellipsis is the conjunctions. In case, the conjunction is recognized as a conjunction of range mentions (C_R state), all values in the range of these two suffixes should be considered as conjunct candidates. Otherwise, once the conjunction is recognized as a conjunction of mentions with coordination ellipsis (C state), the candidates in the conjuncts region are independent to each other, and dependent to the antecedent region. Therefore, the reassembly mentions are the combinations of antecedent region and each conjunct candidate.

As shown in Fig. 3(b), individual and overlap abbreviation pair mentions belong to Pattern 2. The pair (long form "neurokinin-3" and abbreviation "NK-3") of abbreviation mentions is in the conjuncts region. Therefore, the two candidates, long form and abbreviation, are reassembled with antecedent region individually. We detected the long form region by applying the Ab3P abbreviation identification tool [40]. Thus, we are able to identify the conjuncts region in these mentions. Patterns 3 and 4 in Fig. 2 are relatively easier than Pattern 1 and 2. Since the patterns do not contain a conjuncts region, assembling the individual mentions only requires splitting conjunctions and parentheses. As shown in Fig. 3(c)/(d), the mentions can be separated individually.

In addition to the aforementioned types, mixed mentions are more complicated. We defined a two-phase strategy to divide concepts in mixed mentions. In the first phase, we split the mention using Patterns 3 or 4, which do not contain any conjuncts region. In this phase, all conjunctions of range (C_R), mention with coordination ellipsis (C) and abbreviation parentheses (L/R_o) are considered as part of antecedent region. In the second phase, the mention is decomposed by Pattern 1 or 2. In this phase, we start to face the conjuncts region. As shown in Fig. 4, "interferon gamma (IFN-gamma)-inducible protein 10 (gamma IP-10)" is split to "interferon gamma (IFN-gamma)-inducible protein 10" and "gamma IP-10" by cutting in first phase. According the states L/R_o, which have been identified in "interferon gamma

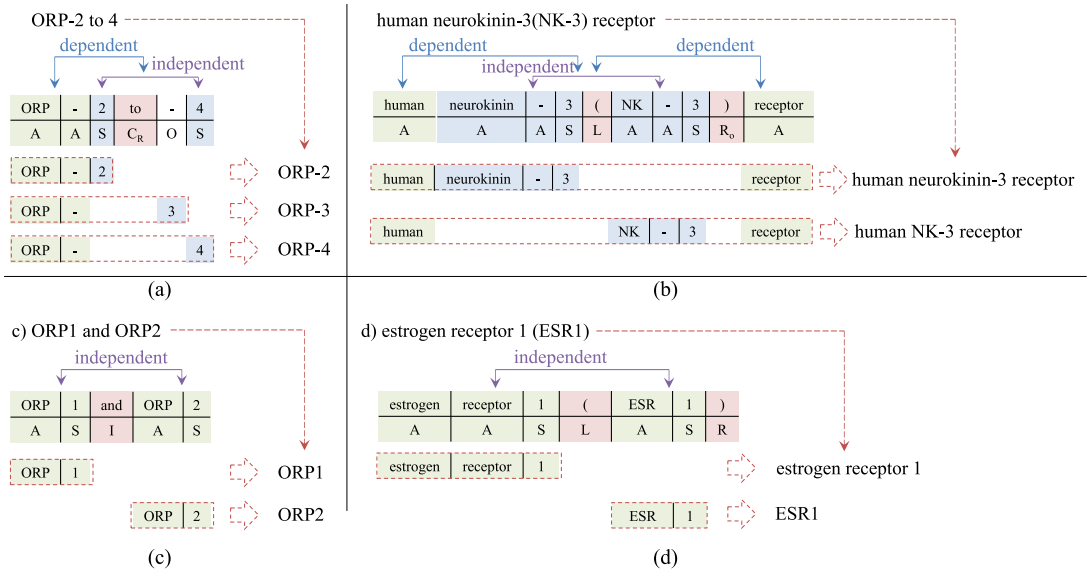


Fig. 3. Strategy of reassembly for mention with coordination ellipsis, range mention, and abbreviation.

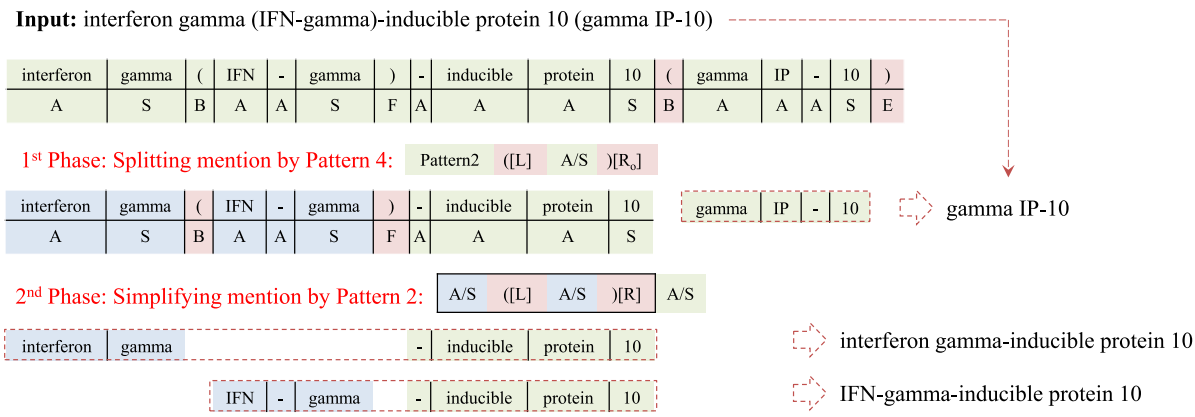


Fig. 4. Strategy of reassembly for mixed mention.

(IFN-gamma)-inducible protein 10," the second phase should choose Patten 2 for simplifying. We, therefore, obtained "interferon gamma-inducible protein 10" and "IFN-gamma-inducible protein 10" from the second phase simplification.

In other words, the main idea of this two-phase strategy is to retain all submentions with a conjuncts region in the second phase. Since the submentions that map to Patterns 1 & 2 are more complicated and cannot be separated individually, those submentions will be processed in the second step.

D. Postprocessing

To optimize the CRF result in SimConcept, we developed several heuristic rules in postprocessing. The first rule is concerned with some plural mentions, such as "SMADs 1 and 3." Typically, we remove the letter "s" from individual mentions when extracting them from the composite mention. For example, the output of "SMADs 1 and 3" becomes "SMAD 1" and "SMAD 3." However, in some cases, the letter "s" is actually part of an entity name in individual mentions (e.g., "Vps 35"). Therefore, we first extract individual mentions without the letter "s" and search for its occurrence in the full text. If we cannot locate it in full text, we will add the "s" suffix to the individual mention. The second postprocessing rule handles some antecedent and prefix tokens that cannot be normally split by our tokenization module, such as "tri- and diorganton." In such a case, we recognize the prefix (i.e., tri-, di-, mono-, hexachloro-, hexabromo-, and hexa-) and change the state of tokens accordingly (e.g., $\frac{\text{diorganton}}{\mathbf{A}}$ to $\frac{\text{di}}{\mathbf{S}}\frac{\text{organton}}{\mathbf{A}}$). As a result, "tri- and diorganton" is decomposed to "triorganton" and "diorganton."

E. SimConcept Corpus

The SimConcept corpus was compiled using five datasets: three for genes, one for diseases, and one for chemicals. For genes, we integrated the BioCreative II gene normalization task training (281 abstracts) and test (262 abstracts) corpora and the 151 GIA test collection (http://ii.nlm.nih.gov/DataSets/

TABLE I
DESCRIPTIVE STATISTICS FOR THE SIMCONCEPT CORPUS

Concept	# of abstracts	Five types of composite mentions					
		All	C_R	C	I	IA	OA
Gene	694	810 (1895)	14 (60)	101 (246)	442 (1089)	253 (534)	41 (107)
Disease	793	1012 (2293)	2 (18)	245 (583)	303 (809)	486 (1045)	52 (123)
Chemical	937	1012 (2944)	99 (505)	201 (771)	496 (1389)	302 (716)	0 (0)

The numbers of composite mentions (of different types) are first listed followed by the numbers of individual mentions after decomposition in parentheses.

TABLE II
STATISTIC OF SIMCONCEPT CORPUS

	Precision	Recall	F-measure
Gene	89.51%	91.35%	90.42%
Disease	87.92%	85.07%	86.47%
Chemical	87.44%	84.71%	86.05%

index.shtml#GIA). In addition, we also collected disease mention corpus from NCBI Disease corpus [16, 41] with 793 abstracts, and sampled Chemical mention corpus from BioCreative IV CHEMDNER task [42] training dataset for 937 abstracts. As shown in Table I, we collected 2 424 abstracts in total. For each article, in addition to the annotations of all described bioconcept mentions, we appended following annotations: 1) the decomposed mentions, such as "BRCA1" and "BRCA2" for "BRCA1/2"; 2) the five types of composite mentions (e.g., "mention with coordination ellipsis"); and 3) the states of tokens ("$\frac{\mathrm{BRCA}}{\mathbf{A}}\ \frac{1}{\mathbf{S}}\ \frac{/}{\mathbf{C}}\ \frac{2}{\mathbf{S}}$"). We used PubTator [43], [44], a web-based annotation tool to annotate the corpus. The distributions of the five composite mention types (C_R: Range mention, C: mention of coordination ellipsis, I: individual mentions, IA: individual abbreviation, and OA: overlap abbreviation.) in Table I are different between the three sets. Chemicals contain significantly more range mentions than either disease or genes, and diseases contain more individual abbreviations than chemicals or genes. The distribution for genes is more even across all types than either diseases or chemicals.

III. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate our method, we used leave-one-out cross validation on the three sets (i.e., gene, disease, and chemical). Table II shows the results of our evaluation, where we see that the overall performance is high for all three entity types.

The chemical corpus includes a type of mentions, which is not addressed by our patterns. It is a joint mention that the second mention uses coreference to indicate the previous mention, such as "3–0-propargylated betulinic acid and its 1,2,3-triazoles." The pronoun "its" represents the previous mention "3–0-propargylated betulinic acid." Therefore, this composite mention contains two individual mentions "3–0-propargylated betulinic acid" and "1,2,3-triazoles of 3–0-propargylated betulinic acid." We have ignored this type of mentions.

TABLE III
EVALUATION OF INDIVIDUAL MENTION TYPES

	Gene	Disease	Chemical
Individual abbreviation	92.05%	84.21%	86.69%
Overlap abbreviation	80.9 %	91.5 %	N/A
Mention with coordination ellipsis	76.35%	80.21%	61.10%
Range mention	91.67%	N/A	94.14%
Individual mention	91.11%	87.13%	87.34%
Mixed mention	81.75%	81.45%	83.84%
All composite mentions	90.42%	86.47%	86.05%

Scores are F-measures.

TABLE IV
SIMCONCEPT CONTRIBUTION ON GENE NORMALIZATION PERFORMANCE

Method	Precision	Recall	F-measure
GenNorm + SimConcept	87.01%	86.13%	86.57%
GenNorm + heuristic rules	86.78%	85.23%	86.00%
GenNorm	86.72%	84.09%	85.38%

TABLE V
SIMCONCEPT CONTRIBUTION ON DISEASE NORMALIZATION PERFORMANCE

Method	Precision	Recall	F-measure
DNorm + SimConcept	80.91%	79.23%	80.06%
DNorm	80.69%	76.85%	78.72%

To assess the performance on each composite mention type, we computed results shown in Table III. There are only two range mentions in the disease set, and we therefore, ignored these. There are also no overlap mentions in the chemical set. Since two exception mentions belong to continuous mention type in chemical corpus, the performance of continuous mention becomes lower.

As mentioned in introduction, this study is aimed at helping bioconcept normalization. We, therefore, applied SimConcept in GenNorm [19] and DNorm [16], and evaluated on the test sets of BioCreative II gene normalization task [45] and NCBI disease corpus [46], respectively (no normalized chemical corpus is available). To avoid training on the test set, the training set for SimConcept excluded the test corpora for GenNorm and DNorm. As shown in Tables IV and V, using SimConcept can further improve the state-of-the-art performance for 1.17% in F-measure (P-value $= 0.02$) for gene normalization and 1.34% in F-measure (P-value $= 0.03$) for disease normalization. We also applied the heuristic rules used in previous gene normalization studies [47, 48] and showed the result in second row of Table IV. Our set of heuristics includes nine rules. Those rules are defined by regular expressions to recognize the conjuncts at the end of the mention (e.g., detecting "1" and "2" in "BRCA1/2") and handle some mentions containing coordination ellipsis and ranges. However, the composite mentions that are not considered in the refinement of heuristic rules cannot be recognized. This comparison shows that using heuristic rules is not as robust

TABLE VI
PERFORMANCE DECREASE WHEN REMOVING FEATURES (EVALUATED ON GENE CORPUS)

	Precision	Recall	F-measure
SimConcept	89.51%	91.35%	90.42%
- Token features	87.41%	89.27%	88.33% (–2.09%)
- Semantic features	88.64%	89.78%	89.21% (–1.21%)
- Full text features	88.54%	90.34%	89.42% (–1.00%)
- Character features	89.03%	90.42%	89.72% (–0.70%)
- Pattern features	89.50%	91.12%	90.30% (–0.12%)
Order 2 → Order 1	87.61%	89.38%	88.49% (–1.93%)

as SimConcept. As also shown in Table V, using heuristic rules raises performance about half as much as SimConcept.

In order to examine the contribution of individual feature types, we performed a feature ablation study where different feature types were removed from the entire set of features one at a time. As shown in Table VI, the largest drop in performance was due to the removal of token features, followed by semantic and character features. The removal of case pattern or contextual features had little effect on final performance. In addition to removing features, we also changed the order of CRF model from order 2 to order 1. The result shows order 2 performs better than order 1.

Despite our best efforts, there are still errors in our decomposition results. We examined a sample set of errors and grouped them into two major categories. The first group is due to incorrect conjunction type detection. For example, "AKR1C1-AKR1C4" is a range mention including AKR1C1, AKR1C2, AKR1C3, and AKR1C4. However, SimConcept incorrectly recognizes this as a mention with coordination ellipsis, thus missing two individual mentions AKR1C2 and AKR1C3. This category accounts for 84% of all errors in SimConcept. The second category is because of the incorrect antecedent/conjuncts region detection. Approximately 12% of SimConcept errors belong to this second group. For example, "cytosolic and mitochondrial serine hydroxymethyltransferase" is a composite mention with coordination ellipsis, which should be decomposed into "cytosolic serine hydroxymethyltransferase" and "mitochondrial serine hydroxymethyltransferase." However, SimConcept recognizes serine as part of the conjuncts region of "mitochondrial serine" by mistake. As a result, the output of first extracted mention becomes "cytosolic hydroxymethyltransferase."

IV. CONCLUSION

In this study, we present SimConcept—a method to handle the task of composite named entity simplification. We integrated a CRF-based method with a pattern identification strategy to systematically decompose the six types of composite mentions. To handle the three most fundamental bioconcepts, we reannotated the composite mentions in five existing corpora for gene (BioCreative 2 GN task train/test corpus and NLM GIA corpus), disease (NCBI disease corpus), and chemicals (BioCreative IV ChemDNER task corpus), and used these to evaluate SimConcept. The results show that SimConcept handles composite mention simplification effectively.

We further used SimConcept to assist the bioconcept normalization task. The results suggest that SimConcept is helpful for improving normalization performance. Our approach should generalize to other entity types in addition to the three concepts that were the focus of this study: genes, diseases, and chemicals. However, the problems of token reassembly step of different concepts are highly diverse. Using a pattern-based method, may not be able to address all issues. In our future work, we will apply statistical methods (e.g., an unsupervised statistical approach [49]) to handle the reassembly issue.

REFERENCES

[1] C.-H. Wei, R. Leaman, and Z. Lu, "SimConcept: A hybrid approach for simplifying composite named entities in biomedicine," in *Proc. ACM Conf. Bioinformat. Comput. Biol. Health Informat.*, Newport Beach, CA, USA, 2014, pp. 138–146.

[2] M. Krallinger, M. Vazquez, F. Leitner, D. Salgado, A. Chatr-aryamontri, A. Winter, L. Perfetto, L. Briganti, L. Licata, M. Iannuccelli, L. Castagnoli, G. Cesareni, M. Tyers, G. Schneider, F. Rinaldi, R. Leaman, G. Gonzalez, S. Matos, S. Kim, W. J. Wilbur, L. Rocha, H. Shatkay, A. V. Tendulkar, S. Agarwal, F. Liu, X. Wang, R. Rak, K. Noto, C. Elkan, Z. Lu, R. I. Dogan, J.-F. Fontaine, M. A. Andrade-Navarro, and A. Valencia, "The protein-protein interaction tasks of biocreative iii: Classification/ranking of articles and linking bio-ontology concepts to full text," *BMC Bioinformatics*, Suppl 8:S3, 2011.

[3] W. A. Baumgartner Jr., Z. Lu, H. L. Johnson, J. G. Caporaso, J. Paquette, A. Lindemann, E. K. White, O. Medvedeva, K. B. Cohen, and L. Hunter, "An integrated approach to concept recognition in biomedical text," in *Proc 2nd BioCreative Challenge Eval. Workshop*, 2007, pp. 257–271.

[4] H. Poon and L. Vanderwende, "Joint inference for knowledge extraction from biomedical literature," presented at the Human Language Technologies Annu. Conf. North American Chapter Association for Computational Linguistics, Los Angeles, CA, USA, 2010.

[5] L. Hunter, Z. Lu, J. Firby, W. A. Baumgartner, H. L. Johnson, P. V. Ogren, and K. B. Cohen, "OpenDMAP: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression," *BMC Bioinformat.*, 9:78, 2008.

[6] S. Bethard, Z. Lu, J. H. Martin, and L. Hunter, "Semantic role labeling for protein transport predicates," *BMC Bioinformat.*, 9:277, 2008.

[7] C. C. Yang, H. Yang, and L. Jiang, "Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media," *ACM Trans. Manage. Inf. Syst.*, vol. 5, no. 1, art. no. 2, Apr. 2014.

[8] R. I. Doğan, A. Névéol, and Z. Lu, "A context-blocks model for identifying clinical relationships in patient records," *BMC Bioinformat.*, Suppl 3:S3, 2011.

[9] J. Li and Z. Lu, "Systematic identification of pharmacogenomics information from clinical trials," *J. Biomed. Informat.*, vol. 45, pp. 870–878, 2012.

[10] Y. Mao, K. Van Auken, D. Li, C. N. Arighi, P. McQuilton, G. T. Hayman, S. Tweedie, M. L. Schaeffer, S. J. F. Laulederkind, S.-J. Wang, J. Gobeill, P. Ruch, A. T. Luu, J.-j. Kim, J.-H. Chiang, Y.-D. Chen, C.-J. Yang, H. Liu, D. Zhu, Y. Li, H. Yu, E. Emadzadeh, G. Gonzalez, J.-M. Chen, H.-J. Dai, and Z. Lu, "Overview of the gene ontology task at BioCreative IV," *Database*, vol. 2014, bau086, 2014.

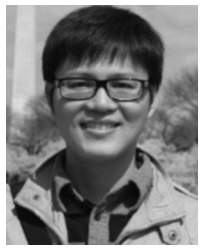
[11] C. N. Arighi, C. H. Wu, K. B. Cohen, L. Hirschman, M. Krallinger, A. Valencia, Z. Lu, J. W. Wilbur, and T. C. Wiegers, "BioCreative-IV virtual issue," *Database*, vol. 2014, bau039, 2014.

[12] Z. Lu, H.-Y. Kao, C.-H. Wei, M. Huang, J. Liu, C.-J. Kuo, C.-N. Hsu, R. T.-H. Tsai, H.-J. Dai, N. Okazaki, H.-C. Cho, M. Gerner, I. Solt, S. Agarwal, F. Liu, D. Vishnyakova, P. Ruch, M. Romacker, F. Rinaldi, S. Bhattacharya, P. Srinivasan, H. Liu, M. Torii, S. Matos, D. Campos, K. Verspoor, K. M. Livingston, and W. J. Wilbur, "The gene normalization task in BioCreative III," *BMC Bioinformat.*, Suppl 8:S2, 2011.

[13] C. N. Arighi, Z. Lu, M. Krallinger, K. B. Cohen, W. J. Wilbur, A. Valencia, L. Hirschman, and C. H. Wu, "Overview of the BioCreative III workshop," *BMC Bioinformat.*, Suppl 8: S1, 2011.

[14] A. Névéol, R. I. Doğan, and Z. Lu, "Semi-automatic semantic annotation of PubMed queries: A study on quality, efficiency, satisfaction," *J. Biomed. Informat.*, vol. 44, pp. 310–318, 2011.

[15] R. I. Dogan, G. C. Murray, A. Névéol, and Z. Lu, "Understanding PubMed user search behavior through log analysis," *Database*, vol. 2009, bap018, 2009.
[16] R. Leaman, R. I. Doğan, and Z. Lu, "DNorm: Disease name normalization with pairwise learning to rank," *Bioinformatics*, vol. 29, pp. 2909–2917, 2013.
[17] C.-H. Wei, H.-Y. Kao, and Z. Lu, "SR4GN: a species recognition software tool for gene normalization," *Plos One*, 7(6): p. e38460, 2012.
[18] T. Rocktäschel, M. Weidlich, and U. Leser, "ChemSpot: A hybrid system for chemical named entity recognition," *Bioinformatics*, vol. 28, pp. 1633–1640, 2012.
[19] C.-H. Wei and H.-Y. Kao, "Cross-species gene normalization by species inference," *BMC Bioinformat.*, vol. 12, S5, 2011.
[20] M. Torii, K. Wagholikar, and H. Liu, "Detecting concept mentions in biomedical text using hidden Markov model: Multiple concept types at once or one at a time?," *J Biomed. Semantics*, 5(1):3, 2014.
[21] R. Leaman, C.-H. Wei, and Z. Lu, "tmChem: a high performance approach for chemical named entity recognition and normalization," *J Cheminformat.*, 7(Suppl 1):S3, 2015.
[22] S. Van Landeghem, J. Björne, C.-H. Wei, K. Hakala, S. Pyysalo, S. Ananiadou, H.-Y. Kao, Z. Lu, T. Salakoski, Y. Van de Peer, and F. Ginter, "Large-scale event extraction from literature with multi-level gene normalization," *PloS One*, 8(4): e55814, 2013.
[23] G. Leroy, J. E. Endicott, O. Mouradi, A. Kauchak, Melissa, and L. Just, "Improving perceived and actual text difficulty for health information consumers using semi-automated methods," presented at the American Medical Informatics Association Annu. Symp., Chicago, IL, USA, 2012.
[24] E. Ong, J. Damay, G. Lojico, K. Lu, and D. Tarantan, "Simplifying text in medical literature," *J. Res. Sci., Comput. Eng.*, vol. 4, pp. 37–47, 2007.
[25] A. Siddharthan, "Syntactic simplification and text cohesion," *Res. Lang. Comput.*, vol. 4, pp. 77–109, 2006.
[26] R. Chandrasekar and B. Srinivas, "Automatic induction of rules for text simplification," *Knowledge-Based Syst.*, vol. 10, pp. 183–190, 1997.
[27] D. Kauchak, "Improving text simplification language modeling using unsimplified text data," presented at the 51st Annu. Meet. Association Computational Linguistics, Sofia, Bulgaria, 2013.
[28] S. B. Silveira and A. Branco, "Enhancing multi-document summaries with sentence simplification," presented at the Int. Conf. Artificial Intelligence, Las Vegas, NV, USA, 2012.
[29] Y. Peng, C. O. Tudor, M. Torii, C. H. Wu, and K. Vijay-Shanker, "Isimp: A sentence simplification system for biomedical text," presented at the IEEE Int. Conf. Bioinformatics Biomedicine, Philadelphia, PA, USA, 2012.
[30] D. Vickrey and D. Koller, "Sentence simplification for semantic role labeling," presented at the 22nd Int. Conf. Computational Linguistics, Stroudsburg, PA, USA, 2008.
[31] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion," *Inf. Proc. Manag.*, vol. 43, pp. 1606–1618, 2007.
[32] E. Buyko, K. Tomanek, and U. Hahn, "Resolution of coordination ellipses in biological named entities using conditional random fields," presented at the 10th Conf. Pacific Association for Computational Linguistics, Melbourne, Australia, 2007.
[33] J.-D. Kim, T. Ohta, Y. Tateisi, and J. I. Tsujii, "GENIA corpus—A semantically annotated corpus for bio-text mining," *Bioinformatics*, vol. 19, pp. i180–i182, 2003.
[34] J. Chae, Y. Jung, T. Lee, S. Jung, C. Huh, G. Kim, and H. Oh, "Identifying non-elliptical entity mentions in a coordinated NP with ellipses," *J. Biomed. Inf.*, pp. 139–152, 2013.
[35] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," presented at the Int. Conf. Machine Learning, Williamstown, MA, USA, 2001.
[36] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program. B*, vol. 45, pp. 503–528, 1989.
[37] C.-H. Wei, B. R. Harris, H.-Y. Kao, and Z. Lu, "tmVar: A text mining approach for extracting sequence variants in biomedical literature," *Bioinformatics*, vol. 29, pp. 1433–1439, 2013.
[38] D. M. Lowe, P. T. Corbett, P. Murray-Rust, and R. C. Glen, "Chemical name to structure: OPSIN, an open source solution," *J. Chem. Inf. Modeling*, vol. 51, pp. 739–753, 2011.
[39] L. Tanabe and W. J. Wilbur, "Tagging gene and protein names in biomedical text," *Bioinformatics*, vol. 18, pp. 1124–1132, 2002.
[40] S. Sohn, D. C. Comeau, W. Kim, and W. J. Wilbur, "Abbreviation definition identification based on automatic precision estimates," *BMC Bioinformatics*, 9:402, 2008.
[41] R. I. Doğan and Z. Lu, "An improved corpus of disease mentions in PubMed citations," presented at the Workshop Biomedical Natural Language Processing, Montreal, Canada, pp. 91–99, 2012.
[42] M. Krallinger, O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D. Ji, D. M. Lowe, R. A. Sayle, R. T. Batista-Navarro, R. Rak, T. Huber, T. Rocktäschel, S. Matos, D. Campos, B. Tang, H. Xu, T. Munkhdalai, K. H. Ryu, S. Ramanan, S. Nathan, S. Žitnik, M. Bajec, L. Weber, M. Irmer, S. A. Akhondi, J. A. Kors, S. Xu, X. An, U. K. Sikdar, A. Ekbal, M. Yoshioka, T. M. Dieb, M. Choi, K. Verspoor, M. Khabsa, C. L. Giles, H. Liu, K. E. Ravikumar, A. Lamurias, F. M. Couto, H.-J. Dai, R. T.-H. Tsai, C. Ata, T. Can, A. Usié, R. Alves, I. Segura-Bedmar, P. Martínez, J. Oyarzabal, and A. Valencia, "The CHEMDNER corpus of chemicals and drugs and its annotation principles," *J. Cheminformat.*, 7(Suppl 1):S2, 2015.
[43] C.-H. Wei, H.-Y. Kao, and Z. Lu, "PubTator: A Web-based text mining tool for assisting Biocuration," *Nucleic Acids Res.*, vol. 41, pp. W518–W522, 2013.
[44] C.-H. Wei, B. R. Harris, D. Li, T. Z. Berardini, E. Huala, H.-Y. Kao, and Z. Lu, "Accelerating literature curation with text-mining tools: A case study of using PubTator to curate genes in PubMed abstracts," *Database: J. Biol. Databases Curation*, vol. 2012, bas041, 2012.
[45] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch *et al.*, "Overview of BioCreative II gene normalization," *Genome Biol.*, vol. 9, p. S3, 2008.
[46] R. I. Doğan, R. Leaman, and Z. Lu, "NCBI disease corpus: A resource for disease name recognition and concept normalization," *J. Biomed. Inf.*, vol. 47, pp. 1–10, Jan. 2014.
[47] J. Wermter, K. Tomanek, and U. Hahn, "High-performance gene name normalization with GeNo," *Bioinformatics*, vol. 25, pp. 815–821, 2009.
[48] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez, "Inter-species normalization of gene mentions with GNAT," *Bioinformatics*, vol. 24, pp. i126–i132, 2008.
[49] C. Zhai, "Fast statistical parsing of noun phrases for document indexing," in *Proc. 5th Conf. Appl. Natural Lang. Proc.*, 1997, pp. 312–319.



Chih-Hsuan Wei received the Ph.D. degree in computer science and information engineering from the National Cheng-Kung University, Tainan, Taiwan.

He is currently a Research Fellow at the National Center for Biotechnology Information, Bethesda, MD, USA, and dedicates on bioconcept normalization and biocuration assistance.

Robert Leaman received the Ph.D. degree in computer science from Arizona State University, Tempe, AZ, USA.

He is currently a research fellow at the National Center for Biotechnology Information, Bethesda, MD, USA. His research focuses on bioconcept entity recognition and normalization.

Zhiyong Lu received the Ph.D. degree in biomedical informatics from the University of Colorado School of Medicine, Aurora, CO, USA.

He is Earl Stadtman Investigator at the National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA, where he leads the biomedical text mining research group. His research has been applied to NCBI's PubMed system and beyond.