Contents

Annotation Guideline for ChEMU-Ref: A Corpus for Modeling Anaphora Resolution in the Chemical Domain

Biaoyan Fang[1], Christian Druckenbrodt[2], Colleen Yeow Hui Shiuan[1], Sacha Novakovic[1], Ralph Hössel[2], Saber A. Akhondi[2], Jiayuan He[1,3], Meladel Mistica[1], Timothy Baldwin[1] and Karin Verspoor[1]

[1]The University of Melbourne, Australia

[2]Elsevier

[3]RMIT University, Australia

1 Overview

In biochemistry, chemical compounds play an important role in pharmaceutical research and can help to save many lives from severe diseases. For chemical compound analysis, the discovery of compounds is usually first presented in chemical patents, which makes patent corpus analysis important for biochemical research (Akhondi et al., 2019; Segler et al., 2018). However, extracting actionable knowledge from corpus data has for some time been recognised as a bottleneck for drug discovery (Gwynne and Heabrer, 2015). To tackle this bottleneck, an information extraction system that automatically decomposes the research results, specifically chemical patents, into structured data, can be useful in facilitating the process of finding, relating, and reasoning for drug discovery. What's more, with the rapid grown of biochemical publications, e.g. over 1.2 million biotechnical articles published in 2015 alone, the scale of the publication is increasing annually (Li et al., 2016). It is challenging for researchers to find related biotechnical documents, and a direct obstacle to research progress. Furthermore, it also affects the development of biochemistry since the exponential growth of literature brings increasing difficulties in categorizing and navigating publications. Thus, it is meaningful to develop a natural language processing (NLP) tool for information extraction system that automatically transforms unstructured data, mainly text data, into structured, queryable data providing actionable information.

To build this kind of information extraction system for biochemical text, especially for chemical patents, one of the most critical tasks is to extract reaction information, including chemical products, reaction conditions, the interaction of different products, etc. However, in natural language text, including biotech literature, there are various referring relationships needed to be concerned among expressions and it is critical in understanding text. For instance, linguistic

"short cuts" (pronouns, abbreviations, etc.) is applied to avoid redundancy in repeating names or complex descriptions (Choi et al., 2016). This is one of the major obstacles that limit the performance of information extraction systems, since systems need to figure out which entity is referred to in a given context (Li et al., 2013; Vanegas et al., 2015). To tackle this problem of anaphora resolution, available annotated corpora are important. Furthermore, to construct annotated corpora, a feasible annotation framework must be developed.

The purpose of this annotation task is to model referring relationship types within various expressions in chemical patents and "link" all the specific expressions that contain referring relationship. As in example (1), there are many referring relationships among various expressions. ⌈**(R)-tert-butyl (1-hydroxy-2-methylhexan-2-yl)carbamate**⌋$_1$, ⌈**1e**⌋$_2$ and ⌈**compound 1e**⌋$_{15}$ have referring relationships, known as coreference (defined in section 2), which means referring to the same entity. The phrase ⌈**the organic solvents**⌋$_6$ has referring relationships called bridging (defined in section 4), which represents context-based referring links, with the previous chemical compounds (i.e. ⌈**a solution of 1d (1 g, 7.6 mmol) in THF (35 mL)**⌋$_3$, ⌈**sat. NaHCO3(aq) (35 mL)**⌋$_4$ and ⌈**di-tert-butyl dicarbonate (3.33 g, 15.24 mmol)**⌋$_5$) as the initial compounds in ⌈**the organic solvents**⌋$_6$ come from previous ones.

(1) Synthesis of ⌈**(R)-tert-butyl (1-hydroxy-2-methylhexan-2-yl)carbamate**⌋$_1$ (⌈**1e**⌋$_2$) To ⌈**a solution of 1d (1 g, 7.6 mmol) in THF (35 mL)**⌋$_3$ was added ⌈**sat. NaHCO3(aq) (35 mL)**⌋$_4$ followed by ⌈**di-tert-butyl dicarbonate (3.33 g, 15.24 mmol)**⌋$_5$. After 24 h, ⌈**the organic solvents**⌋$_6$ were removed in vacuo. ⌈**The resulting slurry**⌋$_7$ was diluted with ⌈**water (50 mL)**⌋$_8$, extracted with ⌈**EtOAc (100 mL)**⌋$_9$, washed with ⌈**brine (10 mL)**⌋$_{10}$, dried over ⌈**Na2SO4**⌋$_{11}$, and concentrated in vacuo. ⌈**The residue**⌋$_{12}$ was subjected to ⌈**silica gel**⌋$_{13}$ chromatography eluting with ⌈**hexanes-EtOAc**⌋$_{14}$ to provide ⌈**compound 1e**⌋$_{15}$.

In the following section, this annotation guideline discusses the referring phenomena and defines referring relationships in chemical patents. More specifically, since the goal of this annotation guideline is to label the referring phenomena in chemical patents, the expressions that are not related to chemical compounds or not involved in referring relationships will not be annotated.

2 Coreference

Coreference is an important type of referring relationship. It is defined as expressions/mentions that refer to the same entity. As demonstrated in example (2), the expressions ⌈**her**⌋$_3$ and ⌈**Queen Elizabeth II**⌋$_2$ refer to the same person and the description ⌈**her**⌋$_9$ stands for ⌈**Princess Margaret**⌋$_5$ in this context. Also, ⌈**a renowned speech therapist**⌋$_7$ and ⌈**Nancy Logue**⌋$_8$ represent the same person. These referring links are labeled as coreference.

(2) ⌈**The Queen Mother**⌋$_1$ asked ⌈**Queen Elizabeth II**⌋$_2$ to transform ⌈⌈**her**⌋$_3$ **sister**⌋$_4$, ⌈**Princess Margaret**⌋$_5$, into ⌈**a variable princess**⌋$_6$ by summoning ⌈**a renowned speech therapist**⌋$_7$, ⌈**Nancy Logue**⌋$_8$, to treat ⌈**her**⌋$_9$ speech impediment.

In chemical patents, coreference plays an important role and helps to track the reaction process. As the example (1) shown above, ⌈**1e**⌋$_2$ is the identifier of ⌈**(R)-tert-butyl (1-hydroxy-2-methylhexan-2-yl)carbamate**⌋$_1$. They stand for the same entity and the referring relationship between them is coreference. It is the same when ⌈**compound 1e**⌋$_{15}$ is introduced later. This referring information is very useful to understand what is discussing in chemical patents.

In general, the referring mention which cannot be interpreted on its own is called *anaphor* and the mention to which it refers back to is called *antecedent*. In this annotation, coreference relationship will be presented using the notation: Coreference (*antecedent*, *anaphor*) and the referring direction is from *anaphor* to its corresponding *antecedent*. E.g. Coreference links from ⌈**the title compound**⌋$_2$ (*anaphor*) to ⌈**(2,6-dichloro-4-fluorophenyl)hydrazine hydrochloride**⌋$_1$ (*antecedent*) in example (3).

(3) Step 1: ⌈**(2,6-dichloro-4-fluorophenyl)hydrazine hydrochloride**⌋$_1$
To a - 5 ^{o}C solution (internal temperature, wet ice/acetone bath) of 2,6-dichloro-4-fluoroaniline (3.0 g, 17 mmol) in 37 % hydrochloric acid (30 mL) and trifluoroacetic acid (20 mL) was added dropwise an aqueous solution of sodium nitrite (1.4 g, 20 mmol, 6 mL water)...The mixture was filtered and the collected solid was washed with isopropyl alcohol and dried under house vacuum to provide ⌈**the title compound**⌋$_2$.

- Coreference (⌈**(2,6-dichloro-4-fluorophenyl)hydrazine hydrochloride**⌋$_1$, ⌈**the title compound**⌋$_2$)

However, there are various types of coreference, which means various expressions can be used to describe the same entity. In the following section, this annotation guideline discusses the coreference types that are concerned in chemical patents.

2.1 Quantified chemical compounds

In chemical patents, products are often described with quantities (e.g. *200 ml of tetrahydrofuran*, *1% NiCl2/CrCl2*, *compound 1(100 mg, 0.32 mmol)*, etc.). It is informative to consider quantity information of compounds when analyzing reaction processes. However, quantity information is treated as an attribute of the compound mentions. For coreference annotation, quantity information will not uniquely identify mentions and mentions with different quantities can be considered as the same entity. As in example (4), ⌈**4.03 g of N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_2$ contains quantity information *4.03 g* and this information is not taken into consideration when annotating coreference. Therefore, ⌈**N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_1$ and ⌈**4.03 g of N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_2$ can be annotated as coreference.

(4) Method of Making ⌈**N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_1$
...The crude product was recrystallized from DMF, yielding ⌈**4.03 g of N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_2$ as white powder.

- Coreference (⌈**N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_1$, ⌈**4.03 g of N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_2$)

(5) Synthesis of ⌈**Compound A-3**⌋$_1$
...The residue was purified by flash column chromatography with dichloromethane-hexane (1:9 to 2:8) to give ⌈**Compound A-3 (0.79 g, 42% yield)**⌋$_2$ as white solid.

- Coreference (⌈**Compound A-3**⌋$_1$, ⌈**Compound A-3 (0.79 g, 42% yield)**⌋$_2$)

(6) (2) ⌈**Methanol**⌋$_1$ Synthesis
200 g of distilled water was added to the methyl bisulfate obtained above and ethanol as the internal standard was added thereto. The reaction was allowed to proceed at 90 ^{o}C. for 4 h. After completion of the reaction, ⌈**the reaction product**⌋$_2$ was analyzed by HPLC. The results are shown in FIG. 2, confirming the production of ⌈**0.51 g of methanol**⌋$_3$.

- Coreference (⌈**Methanol**⌋$_1$, ⌈**the reaction product**⌋$_2$)
- Coreference (⌈**the reaction product**⌋$_2$, ⌈**0.51 g of methanol**⌋$_3$)

A related issue is that even though the quantity of mentions is not taken into consideration during coreference annotation, whether the coreference exists between mentions with different quantities is still based on context. For instance, chemical compounds used in different stages of reaction might not be considered as the same, as shown in example (7). Although the quantities of ⌈**water (4.9 ml)**⌋$_1$ and ⌈**water**⌋$_2$ are not concerned when considering coreference, they cannot be linked as coreference since they do not refer to the same *water* used during the reaction process.

(7) Acetic acid (9.8 ml) and ⌈**water (4.9 ml)**⌋$_1$ were added to the solution of Compound (4) (0.815 g, 1.30 mmol) in THF (4.9 ml). The mixture was stirred for 3 hrs at 50 ^{o}C and then cooled to 0 ^{o}C. 2N-sodium hydroxide aqueous solution was added to the mixture until the pH of the mixture became 9. The mixture was extracted with ethyl acetate for 3 times. The combined organic layer was washed with ⌈**water**⌋$_2$ and saturated aqueous sodium chloride...

Similar cases are in the expression of the crude product. The crude product usually represents the main chemical compound mixing with the other compounds with a small proportion. It is not a pure product under chemistry. As shown in example (8), ⌈**The crude product**⌋$_2$ is not pure thus it can not coreferentially link to the output product ⌈**5-(2,3-difluorophenyl)-3-methyl-3,4-dihydro-1H-pyrimidin-2-one**⌋$_1$.

(8) 190.5: ⌈**5-(2,3-difluorophenyl)-3-methyl-3,4-dihydro-1H-pyrimidin-2-one**⌋$_1$
640 mg (1.9 mmol) of 5-(2,3-difluorophenyl)-1-(4-methoxybenzyl)-3-methyl-3,4-dihydro-1H-pyrimidin-2-one is dissolved in 5 mL of trifluoroacetic acid...⌈**The crude product**⌋$_2$ obtained is purified by silica gel chromatography eluted with a dichloromethane/ethyl acetate mixture 80/20. 85 mg (20%) of 5-(2,3-difluorophenyl)-3-methyl-3,4-dihydro-1H-pyrimidin-2-one is obtained.

2.2 Proper nouns

Proper nouns should be treated as atomic mentions when annotating coreference. In chemical patents, proper nouns might include mentions of constituent entities. These mentions might represent other proper nouns while they should not be annotated if they are nested in proper nouns. As in example (9), nested expression *Dicyclohexyl* from ⌈**N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_1$ and ⌈**4.03 g of N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_2$ are not annotated and can not be linked as coreference. Also, *chloride* from ⌈**2-fluoro-5-iodo-benzoyl chloride**⌋$_1$ and ⌈**524 mg 2-fluoro-5-iodo-benzoyl chloride**⌋$_2$ in example (10) are not annotated for coreference. Detail annotation of proper nouns can be seen in section 3.1.

(9) Method of Making ⌈**N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_1$
...The crude product was recrystallized from DMF, yielding ⌈**4.03 g of N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_2$.

- Coreference (⌈**N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_1$, ⌈**4.03 g of N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_2$)

(10) Synthesis of ⌈**2-fluoro-5-iodo-benzoyl chloride**⌋$_1$
500 mg of 2-fluoro-5-iodobenzoic acid was added to a 50 ml eggplant flask, and then 3 ml of thionyl chloride was added, and heated at 77 ^{o}C for 2 hours. The reaction was monitored by thin layer chromatography (TLC). After the reaction was completed, the mixture was cooled to room temperature and dried by rotary evaporation to remove thionyl chloride to give ⌈**524 mg 2-fluoro-5-iodo-benzoyl chloride**⌋$_2$.

- Coreference (⌈**2-fluoro-5-iodo-benzoyl chloride**⌋$_1$, ⌈**524 mg 2-fluoro-5-iodo-benzoyl chloride**⌋$_2$)

2.3 Identifiers of chemical compounds

Chemical compounds are often abbreviated or marked with identifiers. In that case, coreference is annotated between the identifier and its corresponding chemical compounds.

(11) ⌈**N-methylpyrrolidone**⌋$_1$ (⌈**NMP**⌋$_2$) was stirred for 1 day over CaH2 and finally distilled off...

- Coreference (⌈**N-methylpyrrolidone**⌋$_1$, ⌈**NMP**⌋$_2$)

(12) (1-2) Synthesis of ⌈**N-hydroxy-4-methoxybenzimidamide**⌋$_1$ (⌈**Compound 1-2**⌋$_2$)

- Coreference (⌈**N-hydroxy-4-methoxybenzimidamide**⌋$_1$, ⌈**Compound 1-2**⌋$_2$)

(13) ⌈**1-methyl-4-(4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)benzyl)piperazine**⌋$_1$ (⌈**10**⌋$_2$)
...The residue was purified by flash silica-gel column chromatography to give ⌈**compound 10 (14.6 g, 75.3%)**⌋$_3$ as an earth yellow solid.

- Coreference (⌈**1-methyl-4-(4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)benzyl)piperazine**⌋$_1$, ⌈**10**⌋$_2$)
- Coreference (⌈**10**⌋$_2$, ⌈**compound 10 (14.6 g, 75.3%)**⌋$_3$)

(14) ⌈**(S)-4-amino-N-(1-cyano-2-(3-fluoro-4'-((4-methylpiperazin-1-yl)methyl)-[1,1'-biphenyl]-4-yl)ethyl)tetrahydro-2H-pyran-4-carboxamide**⌋$_1$ (⌈**PZ1101**⌋$_2$)
...The residue was purified by preparative HPLC to afford ⌈**PZ1101 (230 mg, 62.5%)**⌋$_3$ as ⌈**a white solid**⌋$_4$.

- Coreference (⌈**(S)-4-amino-N-(1-cyano-2-(3-fluoro-4'-((4-methylpiperazin-1-yl)methyl)-[1,1'-biphenyl]-4-yl)ethyl)tetrahydro-2H-pyran-4-carboxamide**⌋$_1$, ⌈**PZ1101**⌋$_2$)
- Coreference (⌈**PZ1101**⌋$_2$, ⌈**PZ1101 (230 mg, 62.5%)**⌋$_3$)
- Coreference (⌈**PZ1101 (230 mg, 62.5%)**⌋$_3$, ⌈**a white solid**⌋$_4$)

(15) ⌈**2-((6-(4,4-bis(hydroxymethyl)piperidin-1-yl)-6-oxohexyl)oxy)-3a,4,7,7a-tetrahydro-1H-4,7-epoxyisoindole-1,3(2H)-dione**⌋$_1$ ⌈**5i**⌋$_2$
...The crude product was purified on a silica gel column (2% AcOH, 2-10% MeOH, DCM), to give ⌈**3.223 g (76.3%) of diol 5i**⌋$_3$ as a white solid.

- Coreference (⌈**2-((6-(4,4-bis(hydroxymethyl)piperidin-1-yl)-6-oxohexyl)oxy)-3a,4,7,7a-tetrahydro-1H-4,7-epoxyisoindole-1,3(2H)-dione**⌋$_1$, ⌈**5i**⌋$_2$)
- Coreference (⌈**5i**⌋$_2$, ⌈**3.223 g (76.3%) of diol 5i**⌋$_3$)

2.4 Multiple antecedents

For plural expressions, anaphoric mentions can refer to more than one referent, which means multiple antecedents. As in example (16), the description ⌈**they**⌋$_5$ refers to its corresponding *previous mentions* (i.e. ⌈**1-ethyl-3-(3-dimethylaminopropyl) carbodiimide(EDCl, 860 mg, 4.5 mmol)**⌋$_1$, ⌈**1-hydroxybenzotriazole (HOBt; 860 mg, 4.5 mmol)**⌋$_2$, ⌈**diisopropylethylamine (1.6 mL, 9 mmol)**⌋$_3$ and ⌈**the reaction solution**⌋$_4$). In this annotation, the plural expression links to all of its referents separately.

(16) After ⌈**1-ethyl-3-(3-dimethylaminopropyl) carbodiimide(EDCl, 860 mg, 4.5 mmol)**⌋$_1$, ⌈**1-hydroxybenzotriazole (HOBt; 860 mg, 4.5 mmol)**⌋$_2$, ⌈**diisopropylethylamine (1.6 mL, 9 mmol)**⌋$_3$ were added to ⌈**the reaction solution**⌋$_4$, ⌈**they**⌋$_5$ were stirred at room temperature for 6 hours.

- Coreference (⌈**1-ethyl-3-(3-dimethylaminopropyl) carbodiimide(EDCl, 860 mg, 4.5 mmol)**⌋$_1$, ⌈**they**⌋$_5$)
- Coreference (⌈**1-hydroxybenzotriazole (HOBt; 860 mg, 4.5 mmol)**⌋$_2$, ⌈**they**⌋$_5$)
- Coreference (⌈**diisopropylethylamine (1.6 mL, 9 mmol)**⌋$_3$, ⌈**they**⌋$_5$)
- Coreference (⌈**the reaction solution**⌋$_4$, ⌈**they**⌋$_5$)

2.5 Largest logical span of mentions

Apart from the problem discussed in section 2.2, which means the nested mentions in proper nouns are not annotated, various mentions might be nested in other mentions that are not proper nouns. For instance, the mention *a solution of compound 3 (37.8 g, 44 mmol) in DMF (90 mL)* is not proper noun and contains nested expressions *compound 3 (37.8 g, 44 mmol)* and *DMF (90 mL)*. In that case, coreference annotation cover these overlapping expressions since these expressions can be potentially annotated for corefernce links. However, to eliminate redundancy, the largest logical span of mentions will be considered when there are overlapping mentions that could be linked for the same coreference. As shown in example (17), there are overlapping mentions in the expression *a stirred solution of glucose (50 mg, 0.28 mmol) in anhydrous MeOH (10 mL)*, i.e. ⌈⌈**a stirred solution**⌋$_2$ **of** ⌈⌈**glucose (50 mg, 0.28 mmol)**⌋$_3$ **in** ⌈**anhydrous MeOH (10 mL)**⌋$_4$⌋$_5$⌋$_6$. And the overlapping mentions (⌈**a stirred solution**⌋$_2$, ⌈**glucose (50 mg, 0.28 mmol) in anhydrous MeOH (10 mL)**⌋$_5$, etc.) can be used for the same coreference chain. In that case, ⌈**a stirred solution of glucose (50 mg, 0.28 mmol) in anhydrous MeOH (10 mL)**⌋$_6$ will be annotated for coreference. Another example can be seen in (18), which overlapping mentions are demonstrated and the largest logical one is used in coreference annotation.

(17) ⌈**Cyclen (240 mg, 1.4 mmol)**⌋$_1$ was added to ⌈⌈**a stirred solution**⌋$_2$ **of** ⌈⌈**glucose (50 mg, 0.28 mmol)**⌋$_3$ **in** ⌈**anhydrous MeOH (10 mL)**⌋$_4$⌋$_5$⌋$_6$, and the resulting solution was heated under reflux for 16 hours under N2 atmosphere. Complete consumption of ⌈**the starting materials**⌋$_7$ was confirmed by TLC [stationary phase = C-18 TLC, mobile phase = MeOH : 10% NH4OAc (1:1)].

- Coreference (⌈**Cyclen (240 mg, 1.4 mmol)**⌋$_1$, ⌈**the starting materials**⌋$_7$)
- Coreference (⌈**a stirred solution of glucose (50 mg, 0.28 mmol) in anhydrous MeOH (10 mL)**⌋$_6$, ⌈**the starting materials**⌋$_7$)

(18) ⌈⌈**A solution**⌋$_1$ **of** ⌈⌈**1.1 equivalents of 3,4-dichlorophenylisocyanate**⌋$_2$ **in** ⌈**dry** ⌈**THF (ca. 2 mL per mmol)**⌋$_3$⌋$_4$⌋$_5$⌋$_6$ was added dropwise to ⌈⌈**a mixture**⌋$_7$ **of** ⌈⌈**1 equivalent of the appropriate amino**⌋$_8$ **derivative in** ⌈**dry** ⌈**THF (ca. 2 mL per mmol)**⌋$_9$⌋$_{10}$⌋$_{11}$⌋$_{12}$ and the resulting reaction mixture was stirred at room temperature until ⌈**the starting materials**⌋$_{13}$ was consumed as determined by tlc (ca. 1-2 h).

- Coreference (⌈**A solution of 1.1 equivalents of 3,4-dichlorophenylisocyanate in dry THF (ca. 2 mL per mmol)**⌋$_6$, ⌈**the starting materials**⌋$_{13}$)
- Coreference (⌈**a mixture of 1 equivalent of the appropriate amino derivative in dry THF (ca. 2 mL per mmol)**⌋$_{12}$, ⌈**the starting materials**⌋$_{13}$)

One thing worth noticing is that, to maintain the consistency of mentions that are used for coreference annotation, mentions connected with the conjunction "and" need to be handled carefully. As shown in example (19), *K2CO3 (300 mg, 2.2 mmol) and ethyl 3-bromopropanoate*

(100 mg, 0.61 mmol) is not annotated as a single mention since it cannot represent a mixture of these two based on the context. However, *0.134 g Magnesium (5.54 mmol) and THF (5.5 ml)* in example (20) can be considered as a single mention since the expression represents a mixture of these two.

(19) ⌈**K2CO3 (300 mg, 2.2 mmol)**⌋$_1$ and ⌈**ethyl 3-bromopropanoate (100 mg, 0.61 mmol)**⌋$_2$ were added to ⌈**a stirred solution of cyclen (380 mg, 2.2 mmol) in anhydrous CH3CN (10 mL)**⌋$_3$, and the resulting solution was stirred at room temperature for 16 hours under N2 atmosphere. Complete consumption of ⌈**the starting materials**⌋$_4$ was confirmed by TLC [stationary phase = Basic alumina TLC, mobile phase = CH2Cl2 : MeOH (10:1)].

- Coreference (⌈**K2CO3 (300 mg, 2.2 mmol)**⌋$_1$, ⌈**the starting materials**⌋$_4$)
- Coreference (⌈**ethyl 3-bromopropanoate (100 mg, 0.61 mmol)**⌋$_2$, ⌈**the starting materials**⌋$_4$)
- Coreference (⌈**a stirred solution of cyclen (380 mg, 2.2 mmol) in anhydrous CH3CN (10 mL)**⌋$_3$, ⌈**the starting materials**⌋$_4$)

(20) ⌈**0.134 g Magnesium (5.54 mmol) and THF (5.5 ml)**⌋$_1$ were stirred under an argon atmosphere. 0.87 ml 1-Bromoheptane (4.85 mmol) was added.

2.6 Equipment in chemical patents

Apart from the coreference among chemical compounds, the referring information between chemical equipment is also important for understanding the reaction process. The process to the equipment often equal to the process to the chemical compounds inside that equipment. Thus, the referring relationship within the equipment will be taken into consideration as well. As in example (21) and (22), *flask* and *autoclave* are used for the reaction process and these equipment are essential for tracking the reaction.

(21) Azido ester (S,S)-18 (0.1414 g, 0.5330 mmol) was dissolved in methanol (12 mL) and 10% Pd/C (27 mg, 0.025 mmol) was added. ⌈**The flask**⌋$_1$ was flushed under vacuum and filled with hydrogen three times. A hydrogen-filled balloon was fitted to ⌈**the sealed flask**⌋$_2$ through a needle, and stirring was continued for 24 hat room temperature.

- Coreference (⌈**The flask**⌋$_1$, ⌈**the sealed flask**⌋$_2$)

(22) Under a nitrogen atmosphere in a glove box, Ruthenium complex 1b (1.2 mg, 0.002 mmol) and tetrahydrofuran (2 mL) were added to ⌈**a 300-mL Parr autoclave**⌋$_1$. ⌈**The autoclave**⌋$_2$ was sealed and taken out from the glove box, and was filled with dimethylamine gas (6.8 g, 151 mmol) in dry ice bath.

- Coreference (⌈**a 300-mL Parr autoclave**⌋$_1$, ⌈**The autoclave**⌋$_2$)

3 Mentions

As discussed in section 1, the purpose of this annotation task is to capture the referring phenomena among the mentions related to chemical compounds. Thus, only the mentions that are involved in referring relationships and related to chemical compounds will be annotated. The mentions that are considered for referring links are discussed in the following section and examples are demonstrated in table 1. Specifically, Verbs (e.g. *mix, purify, distill*) and descriptions that refer to event (e.g. *the same process, step 5*) will not be annotated in this corpus.

3.1 Proper Nouns

Proper noun is one of the important parts in chemical patents. Chemical compounds are usually named as proper nouns. In this corpus, proper nouns related to chemical compounds are considered as atomic mentions. What's more, as discussed in section 2.1, quantity information of chemical compounds is important for reaction process. Based on this consideration, if proper nouns are described with quantities, the quantifying description will be included in proper nouns annotation. For instance, in example (23), *N,N'-Dicyclohexyl-1,4-benzenedicarboxamide* is presented with quantity *4.03 g* in this reaction process. This quantifying proper noun ⌈**4.03 g of N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_3$ is considered as an atomic mention. More complex examples can be seen in example (24).

(23) Method of Making ⌈**N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_1$
...The crude product was recrystallized from ⌈**DMF**⌋$_2$, yielding ⌈**4.03 g of N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_3$ as white powder.

Proper nouns: ⌈**N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_1$, ⌈**DMF**⌋$_2$, ⌈**4.03 g of N,N'-Dicyclohexyl-1,4-benzenedicarboxamide**⌋$_3$

(24) ⌈**2-Chloro-4-hydroxy-phenylboronic acid (150 mg, 870.22 μmol)**⌋$_1$, ⌈**2-isopropyl-5-bromoimidazole (82.26 mg, 435.11 μmol)**⌋$_2$, ⌈**1,1'-bis (diphenylphosphine) ferrocene palladium chloride (31.84 mg, 435.11 μmol)**⌋$_3$ and ⌈**sodium carbonate (92.23 mg, 870.22 μmol)**⌋$_4$ were stirred in ⌈**5 ml of tetrahydrofuran**⌋$_5$ and 0.5 ml of water.

Proper nouns: ⌈**2-Chloro-4-hydroxy-phenylboronic acid (150 mg, 870.22 μmol)**⌋$_1$, ⌈**2-isopropyl-5-bromoimidazole (82.26 mg, 435.11 μmol)**⌋$_2$, ⌈**1,1'-bis (diphenylphosphine) ferrocene palladium chloride (31.84 mg, 435.11 μmol)**⌋$_3$, ⌈**sodium carbonate (92.23 mg, 870.22 μmol)**⌋$_4$, ⌈**5 ml of tetrahydrofuran**⌋$_5$

One related issue is that proper nouns in the chemical area sometimes make up by their components, as discussed in section 2.2. These components help to understand the chemical compounds but it is unusual to consider them individually in the compound description. In this corpus, nested proper nouns are not considered as mentions. As in example (25), proper

Type	Example
Proper nouns	*THF*
	DIEA
	DCM
	N-[4-(Benzoxazol-2-yl)-methoxyphenyl]-S-methyl-N'-phenyl-isothiourea
	(2,6-dichloro-4-fluorophenyl)hydrazine hydrochloride
	ethyl 3-(1,4,7,10-tetraazacyclododecan-1-yl)propanoate
Noun phrases	*it*
	they
	this
	compound 5
	formed by-product
	the title compound
	the mixture
	the starting materials
	a pure compound
	0.1 g of anhydrous LiCl
	5.5 mL of cyclohexylamine
	81 % of the raw product
	3,4-dihydro-2H-pyran (0.70 ml, 7.67 mmol)
	Compound (1)(0.350 g, 0.737 mmol)
	2,4-dichloro-6-(6-trifluoromethylpyridin-2-yl)-1,3,5-triazine (5.0 g, 16.9 mmol) in tetrahydrofuran (100 mL)
	a solution of 1.5 equiv. of methyliodide
	a stirred solution of cyclen (380 mg, 2.2 mmol) in anhydrous CH3CN (10 mL)
	a Teflon® flask
	an autoclave
Identifiers	*4*
	5i

Table 1: Chemical compound related mentions

nouns ⌈**N-[4-(benzoxazol-2-yl)-methoxyphenyl]-S-methyl-N'-phenyl-isothiourea**⌋$_1$ and ⌈**1.0 equiv. of N-[4-(benzoxazol-2-yl)-methoxyphenyl]-N'-phenylthiourea**⌋$_2$ make up by similar component, i.e. *4-(benzoxazol-2-yl)-methoxyphenyl*, while this component should not be annotated as atomic mention since it is only part of the description of the proper nouns. It is the same case for *ethyl* in example (26). The nested proper nouns are not annotated as atomic mention.

(25) ⌈**N-[4-(benzoxazol-2-yl)-methoxyphenyl]-S-methyl-N'-phenyl-isothiourea**⌋$_1$ (TS-16b)
A solution of 1.5 equiv. of methyliodide in acetone was added dropwise to an ice-cooled mixture of ⌈**1.0 equiv. of N-[4-(benzoxazol-2-yl)-methoxyphenyl]-N'-phenylthiourea**⌋$_2$ and 1.0 equiv. of potassium carbonate in acetone (ca 5 mL/mmol).

(26) Synthesis of ⌈**ethyl 3-(1,4,7,10-tetraazacyclododecan-1-yl)propanoate**⌋$_1$
K2CO3 (300 mg, 2.2 mmol) and ⌈**ethyl 3-bromopropanoate (100 mg, 0.61 mmol)**⌋$_2$ were added to a stirred solution of cyclen (380 mg, 2.2 mmol) in anhydrous CH3CN (10 mL).

3.2 Noun phrases

Apart from proper nouns, chemical compounds can be presented as noun phrases (NPs). A NP consists of a noun or pronoun and its modifiers and it forms a large part of compound expressions in chemical patents. Since this annotation guideline aims to capture anaphora phenomena over chemical compounds, only NPs that are related to compounds will be considered as mentions. The types of NPs annotated in this guideline are discussed as below.

A pronoun is a word that can function as a NP used by itself and that refers either to the participants in the discourse or to someone or something mentioned elsewhere in the discourse. In chemical patents, a pronoun usually refers to the chemical compound mentioned before. As shown in example (27) and (28), pronouns that are bold stand for chemical compounds and they are considered as mentions.

(27) After 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDCl, 860 mg, 4.5 mmol), 1-hydroxybenzotriazole (HOBt; 860 mg, 4.5 mmol), diisopropylethylamine (1.6 mL, 9 mmol) were added to the reaction solution, ⌈**they**⌋$_1$ were stirred at room temperature for 6 hours. After the reaction was completed, ⌈**it**⌋$_2$ was extracted with ethyl acetate and water.

(28) The compound 1 (398.4 mg, 1.22 mmol) was charged into an eggplant flask, and 5 mL of dichloromethane was then added. ⌈**This**⌋$_1$ was stirred at room temperature for 5 minutes, and then 7 mL of TFA was added, and stirred at room temperature for 2 hours.

Generally, NP contains a noun and its modifiers. A few examples (29) - (31) shown below illustrate NPs that are related to chemical compounds, labelling with bold, and these NPs are considered as mentions.

(29) ⌈**The solvent**⌋$_1$ was evaporated off under reduced pressure, and ⌈**the resulting residue**⌋$_2$ was purified by ⌈**silica gel**⌋$_3$ column chromatography (⌈**hexane-ethyl acetate**⌋$_4$) to obtain ⌈**the title compound**⌋$_5$.

(30) After cooling ⌈**the reaction solution**⌋$_1$ to room temperature, ⌈**5% aqueous solution of hydrochloric acid (30 mL)**⌋$_2$ was added to ⌈**the reaction solution**⌋$_3$ to terminate the reaction. ⌈**The mixture**⌋$_4$ was extracted with ⌈**dichloromethane (30 mL x 3)**⌋$_5$.

(31) ⌈**The reaction mixture**⌋$_1$ was diluted with ⌈**10 ml of methanol**⌋$_2$ and ⌈**10 ml of methylene chloride**⌋$_3$ and concentrated to give ⌈**a crude product**⌋$_4$ which was purified by ⌈**preparative separation plate**⌋$_5$ to give ⌈**a secondary crude product**⌋$_6$ which was then purified by ⌈**prep-HPLC (HCl system)**⌋$_7$ to give ⌈**a pure compound of Example 166 (5.00 mg, the yield was 6.71%)**⌋$_8$ as ⌈**a white powder**⌋$_9$.

What's more, there are a few types of NPs needed to notice and handle carefully in chemical patents: quantified NPs, NPs with preposition and NP with state description.

As discussed in section 2.1, chemical products are usually described with a quantity as it is common that the quantity of products will affect the process of the chemical reaction. NPs with quantities (i.e. quantified NPs) are considered as atomic mentions if the quantities are provided. For instance, ⌈**The compound 1 (398.4 mg, 1.22 mmol)**⌋$_1$ and ⌈**5 mL of dichloromethane**⌋$_2$ in example (32) will be considered as atomic mentions.

One related issue is that quantity expression itself might represent the chemical compound if the chemical compound is not provided. As shown in example (34), ⌈**Yield: 0.15 g (58% of theory)**⌋$_7$ represents chemical compound ⌈**6-[5-Bromo-6-(2,4-dimethoxy-benzylamino)-pyridin-2-ylmethyl]-5-methyl-[1,2,5]oxadiazolo[3,4-b]pyridin-7-ylamine Obtained**⌋$_1$ although that compound is not stated in the expression.

(32) ⌈**The compound 1 (398.4 mg, 1.22 mmol)**⌋$_1$ was charged into an eggplant flask, and ⌈**5 mL of dichloromethane**⌋$_2$ was then added.

(33) ⌈**The residue**⌋$_1$ was washed with ⌈**acetone**⌋$_2$ and ⌈**ether**⌋$_3$ to obtain ⌈**the target compound (1.1 g, 76%)**⌋$_4$.

(34) ⌈**6-[5-Bromo-6-(2,4-dimethoxy-benzylamino)-pyridin-2-ylmethyl]-5-methyl-[1,2,5]oxadiazolo[3,4-b]pyridin-7-ylamine Obtained**⌋$_1$ by starting from ⌈**6-(5-Bromo-6-fluoro-pyridin-2-ylmethyl)-5-methyl-[1,2,5]oxadiazolo[3,4-b]pyridin-7-ylamine**⌋$_2$ (example 22) and ⌈**2,4-dimethoxy-benzylamine**⌋$_3$ using ⌈**diisopropylethylamine**⌋$_4$ instead of ⌈**potassium fluoride**⌋$_5$. Stirred for 18 hours at 120^{o}C. and purified by RP-HPLC (modifier: ⌈**trifluoroacetic acid**⌋$_6$).
⌈**Yield: 0.15 g (58% of theory)**⌋$_7$.

A NP with preposition considers the NPs connected with preposition (e.g. *in, with, of*). Chemical compounds can be represented in this form as well and it is also considered as mention. As in example (35), the phrase ⌈**2,4-dichloro-6-(6-trifluoromethylpyridin-2-yl)-1,3,5-**

triazine (5.0 g, 16.9 mmol) in tetrahydrofuran (100 mL)⌋$_1$ describes a solvent that contains *2,4-dichloro-6-(6-trifluoromethylpyridin-2-yl)-1,3,5-triazine (5.0 g, 16.9 mmol)* and *tetrahydrofuran (100 mL)* and this phrase is considered as a mention. Another example can be seen in (36).

(35) ⌈**2,4-dichloro-6-(6-trifluoromethylpyridin-2-yl)-1,3,5-triazine (5.0 g, 16.9 mmol) in tetrahydrofuran (100 mL)**⌋$_1$ were added 4-amino-2-trifluoromethyl pyridine (3.3 g, 20.3 mmol) and sodium bicarbonate (2.14 g, 25.3 mmol).

(36) ⌈**A solution of 1.1 equivalents of 3,4-dichlorophenylisocyanate in dry THF (ca. 2 mL per mmol)**⌋$_1$ was added dropwise to ⌈**a mixture of 1 equivalent of the appropriate amino derivative in dry THF (ca. 2 mL per mmol)**⌋$_2$.

In chemistry, diffident states of the chemical compound can distinguish different properties. In that case, NP with state information (e.g. *saturated*, *aqueous*, *at O ^oC*, etc.) is considered as a mention. Expressions ⌈**saturated aqueous ammonium chloride solution (30 mL)**⌋$_2$, ⌈**anhydrous tetrahydrofuran (100 mL)**⌋$_2$, ⌈**a stirred solution of 2-(4-(methoxymethoxy)phenyl)-5,5-dimethyl-1,3-dioxane (1.5 g, 5.95 mmol) in THF (15 mL) at -78 ^{o}C**⌋$_1$ and ⌈**isobutylene (8g, 142.6mmol, condensed at -78C)**⌋$_5$ in example (37) - (40), respectively, are considered as mentions.

(37) ...⌈**Water (30 mL)**⌋$_1$ and ⌈**saturated aqueous ammonium chloride solution (30 mL)**⌋$_2$ were added thereto, and ⌈**the mixture**⌋$_3$ was extracted with ⌈**ethyl acetate (100 mL)**⌋$_4$...

(38) ...⌈**2,3-Dichloro-4-nitropyrimidine (5.8 g, 30 mmol)**⌋$_1$ was dissolved in ⌈**anhydrous tetrahydrofuran (100 mL)**⌋$_2$ and ⌈**the reaction solution**⌋$_3$ was cooled to -78^{o}C...

(39) ...To ⌈**a stirred solution of 2-(4-(methoxymethoxy)phenyl)-5,5-dimethyl-1,3-dioxane (1.5 g, 5.95 mmol) in THF (15 mL) at -78 ^{o}C**⌋$_1$ was added ⌈**n-BuLi (9 ml, 8.9 mmol, 1.5 eq)**⌋$_2$...

Apart from NPs that are directly related to chemical compounds, NPs that describes chemical equipment are also related to chemical compounds, as the process to the equipment equals to the process to the chemical compounds inside. The equipment is also considered as mentions. As shown in example (40), the equipment ⌈**a Teflon® flask**⌋$_3$ and ⌈**The flask**⌋$_4$ are also annotated as mentions.

(40) ⌈**Sulfuric acid (conc, 1ml)**⌋$_1$ was added to ⌈**a solution of D-(R)-4-hydroxyphenylglycine (1.0g, 6.0mmol) in 1,4-dioxane (8ml)**⌋$_2$ placed in ⌈**a Teflon® flask**⌋$_3$. ⌈**The flask**⌋$_4$ was cooled to -78C and ⌈**isobutylene (8g, 142.6mmol, condensed at -78C)**⌋$_5$ was added.

Another related issue is that the distinguish between expressions for chemical compounds and reaction processes need to be handled carefully. For example, *the reaction* in example (41) represents the reaction mixture. However, in example (42), it represents the reaction process. Similar

cases are in the expression for *example*. In example (43), *Example 56A (3.2 g, 32.6 mmol)* stands for the output product from the process Example 56A while *Example 56B* represents the label for reaction process to the chemical compound ⌈**bicyclo[1.1.1]pentane-1-carbaldehyde**⌋$_1$. Another example can be seen in (44). *Example 17A*, *Ex. 16A* and *Ex. 12A* describes the reaction process labels.

(41) ...⌈**The mixture**⌋$_1$ was degassed and ⌈**[(2-di-cyclohexylphosphino-3,6-dimethoxy-2',4',6'-triisopropyl-1,1'-biphenyl)-2-(2'-amino-1,1'-biphenyl)]palladium(II) methanesulfonate methanesulfonate (0.039 g, 0.04 mmol)**⌋$_2$ was added. ⌈**The reaction**⌋$_3$ was heated to 100^{o}C. for 3 hours. ⌈**The reaction mixture**⌋$_4$ was diluted with ⌈**DCM (50 mL)**⌋$_5$ and washed with ⌈**water (50 mL)**⌋$_6$...

(42) ...An additional portion of ⌈**sodium borohydride (10 mg, 0.27 mmol)**⌋$_1$ was added, and ⌈**the reaction mixture**⌋$_2$ was heated at 50^{o}C. for an additional 2 h. The reaction was quenched with ⌈**water**⌋$_3$, and ⌈**the reaction mixture**⌋$_4$ was partitioned between ⌈**sat. NaHCO3**⌋$_5$ and ⌈**EtOAc**⌋$_6$...

(43) Example 56B ⌈**bicyclo[1.1.1]pentane-1-carbaldehyde**⌋$_1$
⌈**Example 56A (3.2 g, 32.6 mmol)**⌋$_2$ was dissolved in ⌈**5 mL of dichloromethane**⌋$_3$...

(44) Example 17A ⌈**3-Ethyl-1-(2-methoxyethyl)-5-methyl-2,4-dioxo-1,2,3,4-tetrahydrothieno[2,3-d]pyrimidine-6-carboxylic acid**⌋$_1$
Analogously to the method described in Ex. 16A, ⌈**2.50 g (6.78 mmol) of the compound from Ex. 12A**⌋$_2$ were used to obtain ⌈**1.82 g (85% of theory) of the title compound**⌋$_3$. The reaction time in this case was 1 h.

Also, in this annotation, we tends to leave out the reference information of the chemical compound unless it is necessary to identify the chemical compound. As showed in example (45) and (46), reference information, i.e. *see international patent application WO 2003057673* and *table I*, is not included for the mention annotation. It is different from mention ⌈**2.50 g (6.78 mmol) of the compound from Ex. 12A**⌋$_2$ in example (44), which needs reference information to indicate the chemical compound.

(45) ...⌈**A mixture of 3,5-dimethyl-4-nitro-1 H-pyrazole**$_1$ (see international patent application WO 2003057673) **(3.982 g, 28.22 mmol), iodobenzene (11.3 mL, 101.4 mmol), K2CO3 (8.19 g, 59.26 mmol), CuI (268.8 mg, 1.41 mmol), and trans-N,N'-di-methylcyclohexane-1,2-diamine (890 μL, 5.63 mmol)**⌋$_1$ were heated to 180 ^{o}C for 4 h...

(46) Example 1: ⌈**compound (2)**⌋$_1$ in table I
⌈**4-chloro-3-nitrophenol (5 g, 28.8 mmol, 1 eq.)**⌋$_2$ was placed in ⌈**dimethylformamide (96 mL)**⌋$_3$...

3.3 Identifiers

In chemical patents, identifiers or labels may also be used to represent chemical compounds, in the form of numbers and/or letters. These identifiers are treated as atomic mentions as well.

(47) ($\lceil$**4**$\rfloor_1$) $\rightarrow$ ($\lceil$**5**$\rfloor_2$)
Acetic acid (9.8 ml) and water (4.9 ml) were added to the solution of $\lceil$**Compound (4) (0.815 g, 1.30 mmol)**$\rfloor_3$ in THF (4.9 ml)...

- example of identifiers that represent compounds: $\lceil$**4**$\rfloor_1$, $\lceil$**Compound (4) (0.815 g, 1.30 mmol)**$\rfloor_3$

(48) Preparation Example 1 (Compound Nos.: $\lceil$**2**$\rfloor_1$ and $\lceil$**6**$\rfloor_2$)

- example of identifiers that represent compounds: $\lceil$**2**$\rfloor_1$, $\lceil$**6**$\rfloor_2$

(49) To the mixture of $\lceil$**6 (15 mg, 26 μmol)**$\rfloor_1$ and $\lceil$**10 (11 mg, 40 μmol)**$\rfloor_2$ in a solvent system of methanol-H2O-CH2Cl2 (1:1:1) were added CuSO4·5H2O (6 mg, 26 μmol) and sodium ascorbate (5 mg, 26 μmol) and the mixture was stirred overnight at room temperature.

- example of numbers that represent compounds: $\lceil$**6 (15 mg, 26 μmol)**$\rfloor_1$, $\lceil$**10 (11 mg, 40 μmol)**$\rfloor_2$

However, one related issue is that only the numbers that represent chemical compounds are considered as mentions. The numbers that represent reaction condition, e.g. pH, temperature and time, are not annotated as mentions.

What's more, as stated in Section 3.2, expressions can be used to represent a chemical compound or reaction process. Identifiers also need to be handled carefully. Expression of number with dash or dot can be used to represent a reaction process. As in example (50), expression *2-3* represents "Example 2 Step 3", which is an indicator for a reaction process. Similar example can be seen in example (51).

(50) 2-3 : Preparation of $\lceil$**1,1'-(3-methoxypyrazin-2,6-diyl)diethanone**$\rfloor_1$ ($\lceil$**XI**$\rfloor_2$)
$\lceil$**3,5-Dibromo-2-methoxypyrazine (600 mg, 2.24 mmol)**$\rfloor_3$ and $\lceil$**palladium(II) acetate (40.22 mg, 179.16 μmole)**$\rfloor_4$ were dissolved in...

(51) (1.2) Preparation of $\lceil$**Compound 1B**$\rfloor_1$
$\lceil$**Compound 1A (80.6 g, 400.49 mmol)**$\rfloor_2$, $\lceil$**benzimidamide hydrochloride (69.9 g, 440.61 mmol)**$\rfloor_3$, and $\lceil$**sodium ethoxide**$\rfloor_4$ were put into...

4 Bridging

Apart from coreference, other kinds of anaphora referring expression can be seen in chemical patents. Chemical reaction texts may refer to a reaction product, or a mixture of multiple

chemicals, etc. These might be some intermediate compounds during the reaction process and not introduced explicitly. The relationship between direct chemical mentions and these transformed or created entities cannot be considered to be coreference as they do not refer to the same entity. To tackle this problem, bridging is introduced.

Bridging is defined as capturing semantic relation within mentions and links them via lexicon-semantic, frame, or encyclopedic relations. Specifically, in chemical patents, the semantic relationship of bridging can subsum four types (definition can be seen below): "Transformed", "Reaction-assoicaited", "Work-up" and "Contained".

As the example shows in (52), ⌈**the solution**⌋$_3$ should refer to the solution mixing ⌈**The compound 26 (500 mg, 3.46 mmol)**⌋$_1$ and ⌈**dichloromethane (5 mL)**⌋$_2$. This relationship between these mentions is not coreference since the chemical property of the solution changes when mixing, i.e. some reactions happen. However, this reference is still anaphorically associated since the solution comes from the combination of ⌈**The compound 26 (500 mg, 3.46 mmol)**⌋$_1$ and ⌈**dichloromethane (5 mL)**⌋$_2$ and that referring relationship is considered as "Reaction-associated". Also, as in example (53), the pH of ⌈**the mixture**⌋$_1$ and ⌈**the mixture**⌋$_2$ are different and they are not the same chemical product (entity) in chemistry. Their referring relationship is annotated as "Transformed".

(52) ⌈**The compound 26 (500 mg, 3.46 mmol)**⌋$_1$ was dissolved in ⌈**dichloromethane (5 mL)**⌋$_2$. DIEA (904 μL, 5.19 mmol) and benzyl bromide(494 μL, 4.15 mmol) were added to ⌈**the solution**⌋$_3$.

- Reaction-associated (⌈**The compound 26 (500 mg, 3.46 mmol)**⌋$_1$, ⌈**the solution**⌋$_3$)
- Reaction-associated (⌈**dichloromethane (5 mL)**⌋$_2$, ⌈**the solution**⌋$_3$)

(53) 2N-sodium hydroxide aqueous solution was added to ⌈**the mixture**⌋$_1$ until the pH of ⌈**the mixture**⌋$_2$ became 9.

- Transformed (⌈**the mixture**⌋$_1$, ⌈**the mixture**⌋$_2$)

For the bridging annotation, it only considers the mentions discussed in section 3. The definition of each subtype of bridging is introduced in the following section and only these four types will be considered as bridging in the corpus. Similar to the annotation of coreference in section 2, bridging relationship will be presented using the same notation: Bridging (*antecedent*, *anaphor*), which the referring direction is from *anaphor* to its corresponding *antecedent* (e.g. Transformed (*antecedent*, *anaphor*)) and the same strategy will be applied to handle various referring relationships (i.e. quantifying mentions, multiple antecedents, overlapping mentions discussed in section 2) in bridging annotation.

4.1 Transformed

Based on the chemical concern, the state change of chemical compounds, i.e. implicit reaction happens when processing compounds, will lead to the change of the identification of compounds.

These compounds under different states can not be linked as coreference while they are still anaphoric associated. To capture this associated relationship, "Transformed" is employed.

"Transformed" is defined as an anaphora link for a set of chemical compounds that are initially based on the same components, which have undergone possible changes through various conditions, which may include pH, temperature, etc. It is one-to-one from the anaphor to the corresponding compound (antecedent) that has the same components. For instance, in example (54), the states of the ⌈**mixture**⌋$_1$ and ⌈**the mixture**⌋$_2$ are different since there is an action "stir". Based on chemical knowledge, the implicit reaction happens when taking "stir" action and the property of ⌈**mixture**⌋$_1$ has changed after that. Thus, these two mentions cannot be linked as coreference. There is the same case for ⌈**the mixture**⌋$_1$ and ⌈**it**⌋$_2$ in example (55). In the chemical area, these mentions are not the same entities since the state has changed. However, these mentions discussed above consist of the same components and are still anaphoric associated. To capture this information, these mentions are linked as "transformed" relation, as they have the same components but in different states.

(54) ⌈**The mixture**⌋$_1$ was stirred at room temperature for 1 day. A 2 mol/L aqueous solution of hydrochloric acid was added to ⌈**the mixture**⌋$_2$.

- Transformed (⌈**the mixture**⌋$_1$, ⌈**the mixture**⌋$_2$)

(55) After stirring for 2 h at 0 ^{o}C, ⌈**the mixture**⌋$_1$ was warmed to room temperature and stirred at this temperature until the starting material was completely consumed (tlc monitoring). Then ⌈**it**⌋$_2$ was filtered and evaporated.

- Transformed (⌈**the mixture**⌋$_1$, ⌈**it**⌋$_2$)

4.2 Reaction-associated

Apart from the state changes of chemical compounds, another issue is that reaction is also needed to be considered when mixing chemical compounds. The mixing process might create a new or intermediate compound which has not introduced yet. The created chemical compound also can not be considered as coreference to the previous chemical since they are not the same entity. In this annotation guideline, "Reaction-associated" is introduced to solve that issue.

"Reaction-associated" is defined as the relationship between a chemical compound and its immediate sources via a mixing process. The immediate sources don't have to be the reagent while they need to end up in the corresponding output. The source compounds retain their original chemical structure. It is one-to-many from the anaphor to the source compounds (antecedents). As in example (56), ⌈**the mixture**⌋$_4$ and the product set (i.e. ⌈**Acetic acid (9.8 ml)**⌋$_1$, ⌈**water (4.9 ml)**⌋$_2$, ⌈**the solution of Compound (4) (0.815 g, 1.30 mmol) in THF (4.9 ml)**⌋$_3$) can not be categorized as set-of or part-of relation since reaction has already happened when mixing products. ⌈**the mixture**⌋$_4$ here coreferential refers to the output product from the reaction of mixing process. In that case, "Reaction-associated" is used to represent the referring relationship between them and helps in capturing the original compounds inforamtion of ⌈**the**

mixture⌋$_4$. Another example can be seen in (57). ⌈**the solution**⌋$_3$ refers to *a solvent* that mixing ⌈**3,4-dihydro-2H-pyran (0.70 ml, 7.67 mmol)**⌋$_1$ and ⌈**the solution of Compound (1) (0.350 g, 0.737 mmol) in anhydrous dichloromethane (10 ml)**⌋$_2$. This referring relationship is annotated as "reaction-associated".

(56) ⌈**Acetic acid (9.8 ml)**⌋$_1$ and ⌈**water (4.9 ml)**⌋$_2$ were added to ⌈**the solution of Compound (4) (0.815 g, 1.30 mmol) in THF (4.9 ml)**⌋$_3$. ⌈**The mixture**⌋$_4$ was stirred for 3 hrs at 50C and then cooled to 0 ^{o}C.

- Reaction-associated (⌈**Acetic acid (9.8 ml)**⌋$_1$, ⌈**The mixture**⌋$_4$)
- Reaction-associated (⌈**water (4.9 ml)**⌋$_2$, ⌈**The mixture**⌋$_4$)
- Reaction-associated (⌈**the solution of Compound (4) (0.815 g, 1.30 mmol) in THF (4.9 ml)**⌋$_3$, ⌈**The mixture**⌋$_4$)

(57) ⌈**3,4-dihydro-2H-pyran (0.70 ml, 7.67 mmol)**⌋$_1$ was added to ⌈**the solution of Compound (1) (0.350 g, 0.737 mmol) in anhydrous dichloromethane (10 ml)**⌋$_2$. To ⌈**the solution**⌋$_3$, camphor sulfonic acid (7 mg, 0.03 mmol) was added.

- Reaction-associated (⌈**3,4-dihydro-2H-pyran (0.70 ml, 7.67 mmol)**⌋$_1$, ⌈**the solution**⌋$_3$)
- Reaction-associated (⌈**the solution of Compound (1)(0.350 g, 0.737 mmol) in anhydrous dichloromethane (10 ml)**⌋$_2$, ⌈**the solution**⌋$_3$)

As we stated, it is not necessary that the chemical compounds is reagents. ⌈**nitrogen atmosphere**⌋$_5$ in example (58) is involved in the reaction process and contributes to the output product ⌈**the mixture**⌋$_7$. Here we also consider ⌈**nitrogen atmosphere**⌋$_5$ is linked with"reaction-associated" relation. Another example can be seen in (59), chemical compound ⌈**5 mL of N,N-dimethylformamides**⌋$_4$ is involved in reaction process while not served as reagent - still linked as "reaction-associated".

(58) To ⌈**Compound 147 (220 mg, 0.411 mmol) in DMF (2 mL) solution**⌋$_1$ were added ⌈**cesium carbonate (268 mg, 0.823 mmol)**⌋$_2$ and ⌈**methyl-2-chloro-2,2-difluoro acetate (0.086 mL, 0.823 mmol)**⌋$_3$, ⌈**the mixture**⌋$_4$ was stirred under ⌈**nitrogen atmosphere**⌋$_5$ at 60^{o}C for 15 minutes. ⌈**Water**⌋$_6$ was added thereto, and ⌈**the mixture**⌋$_7$ was extracted with ⌈**ethyl acetate**⌋$_8$...

- ...
- Reaction-associated (⌈**nitrogen atmosphere**⌋$_5$, ⌈**the mixture**⌋$_7$)

(59) ...Adding ⌈**3-methyl-7-(2-butyne-1-yl)-8-bromoxanthine (0.71 g, 2.4 mmol)**⌋$_1$, ⌈**potassium carbonate (0.53 g, 3.8 mmol)**⌋$_2$ and ⌈**2-bromomethyl-3-cyano-pyrazine (0.52 g, 2.6 mmol)**⌋$_3$. Adding ⌈**5 mL of N,N-dimethylformamides**⌋$_4$ in, heating to 80^{o}C and stirring for 5 hours; after the reaction was completed, pouring ⌈**the reaction liquid**⌋$_5$ into...

- ...
- Reaction-associated (⌈**5 mL of N,N-dimethylformamides**⌋$_4$, ⌈**the reaction liquid**⌋$_5$)

4.3 Work-up

In chemical patents, there might be series of manipulations required and plenty of chemical compounds used to isolate and purify chemical products. These chemical compounds do not take part in reaction while are critical for the reaction process. To represent the relationship within products and the compounds used for that purpose, "Work-up" is introduced.

"Work-up" is defined as the relationship between chemical compounds that used for isolation or purification purpose and their corresponding output products. It is one-to-many from the anaphor to the compound/s (antecedent/s) that are used for work-up process. As in example (60), ⌈**The combined organic layer**⌋$_3$ is extracted from ⌈**The mixture**⌋$_1$ by adding ⌈**ethyl acetate**⌋$_2$. These are not the same entity and coreference therefore is not suitable too capture that referring information. The interpretation of ⌈**The combined organic layer**⌋$_3$ still depends on ⌈**The mixture**⌋$_1$ and ⌈**ethyl acetate**⌋$_2$ as it is extracted from them. In that case, this anaphoric relationship is annotated as "Work-up". Another example is shown in (61).

(60) ⌈**The mixture**⌋$_1$ was extracted with ⌈**ethyl acetate**⌋$_2$ for 3 times. ⌈**The combined organic layer**⌋$_3$ was washed with water and saturated aqueous sodium chloride.

- Work-up (⌈**the mixture**⌋$_1$, ⌈**The combined organic layer**⌋$_3$)
- Work-up (⌈**ethyl acetate**⌋$_2$, ⌈**The combined organic layer**⌋$_3$)

(61) ⌈**The residue**⌋$_1$ was taken up in ⌈**ethyl acetate**⌋$_2$, washed with ⌈**water**⌋$_3$ and ⌈**brine**⌋$_4$, dried over anhydrous sodium sulfate and the solvent was removed in vacuo to yield ⌈**81 % of the raw product**⌋$_5$.

- Work-up (⌈**The residue**⌋$_1$, ⌈**81 % of the raw product**⌋$_5$)
- Work-up (⌈**ethyl acetate**⌋$_2$, ⌈**81 % of the raw product**⌋$_5$)
- Work-up (⌈**water**⌋$_3$, ⌈**81 % of the raw product**⌋$_5$)
- Work-up (⌈**brine**⌋$_4$, ⌈**81 % of the raw product**⌋$_5$)

There can be several steps of work up process during the reaction. If the intermediate product during the work up process is not mentioned explicitly, the work up materials can be linked with the output in the following step. As in example (62), there are two steps of work up process, (i.e. *wash* and *purify*) and no intermediate product among these two steps. Thus, the related work up materials are linked with the output in the *purify* process, i.e. ⌈**the title compound (360 mg, 1.05 mmol, 32%)**⌋$_5$. Similar cases can be seen in example (63).

(62) ...⌈**The precipitate**⌋$_1$ was collected and washed with ⌈**water (2x10 mL)**⌋$_2$. Purification (⌈**SiO2**⌋$_3$, ⌈**n-hexane:dichloromethane (2:1)**⌋$_4$) afforded ⌈**the title compound (360 mg, 1.05 mmol, 32%)**⌋$_5$.

- Work-up (⌈**The precipitate**⌋$_1$, ⌈**the title compound (360 mg, 1.05 mmol, 32%)**⌋$_5$)
- Work-up (⌈**water (2x10 mL)**⌋$_2$, ⌈**the title compound (360 mg, 1.05 mmol, 32%)**⌋$_5$)
- Work-up (⌈**SiO2**⌋$_3$, ⌈**the title compound (360 mg, 1.05 mmol, 32%)**⌋$_5$)
- Work-up (⌈**n-hexane:dichloromethane (2:1)**⌋$_4$, ⌈**the title compound (360 mg, 1.05 mmol, 32%)**⌋$_5$)

(63) ⌈**The reaction mixture**⌋$_1$ is stirred at room temperature for 30 min and then concentrated to dryness under vacuum, diluted with ⌈**dichloromethane**⌋$_2$, and finally hydrolysed with ⌈**a saturated aqueous solution of sodium hydrogen carbonate**⌋$_3$. ⌈**The product**⌋$_4$ is extracted with ⌈**dichloromethane**⌋$_5$.

- Work-up (⌈**The reaction mixture**⌋$_1$, ⌈**The product**⌋$_4$)
- Work-up (⌈**dichloromethane**⌋$_2$, ⌈**The product**⌋$_4$)
- Work-up (⌈**a saturated aqueous solution of sodium hydrogen carbonate**⌋$_3$, ⌈**The product**⌋$_4$)

One thing worth mentioning is that work up process also includes conjugate reforming of the chemical compound. As demonstrated in example (64), ⌈**The reaction mixture**⌋$_1$ and ⌈**The product**⌋$_3$ are conjugate and they are linked as "work-up" relation. The involved material (i.e. ⌈**4 ml of 1N aqueous solution of acetic acid**⌋$_2$) is linked as "work-up" as well.

(64) 2.3 ml (2.3 mmol) of a 1N aqueous solution of lithium hydroxide is added to a solution of 475 mg (1.5 mmol) of [5-(2,3-difluorophenyl)-3-methyl-2,4-dioxo-3,4-dihydro-2H-pyrimidin-1-yl]-methyl acetate in 15 ml of tetrahydrofuran and 3 mL of water. ⌈**The reaction mixture**⌋$_1$ is stirred at room temperature for 2 hours, and then adjusted to pH6 by adding ⌈**4 ml of 1N aqueous solution of acetic acid**⌋$_2$. ⌈**The product**⌋$_3$ is extracted with ethyl acetate.

- Work-up (⌈**The reaction mixture**⌋$_1$, ⌈**The product**⌋$_3$)
- Work-up (⌈**4 ml of 1N aqueous solution of acetic acid**⌋$_2$, ⌈**The product**⌋$_3$)

(65) 6-(2-Amino-6-morpholinopyrimidin-4-yl)-3'-fluoro-5-methoxy-[2,4'-bipyridin]-2'-amine (120 mg, 301.95 μmol) and pyridine hydrochloride (Pyridine HCl) (523.41 mg, 4.53 mmol) were stirred in a sealed tube at 170 ^{o}C for 30 min. ⌈**The resulting mixture**⌋$_1$ was cooled to room temperature, neutralized with ⌈**2 N NaOH solution**⌋$_2$ to provide ⌈**a solid**⌋$_3$.

- Work-up (⌈**The resulting mixture**⌋$_1$, ⌈**a solid**⌋$_3$)
- Work-up (⌈**2 N NaOH solution**⌋$_2$, ⌈**a solid**⌋$_3$)

4.4 Contained

Furthermore, the equipment is an essential part of the reaction process. In chemical patents, compounds are placed in equipment and the process for the equipment equals to the process for the chemical compounds inside that equipment. The process for the equipment is also informative. In this case, the previous bridging relations are not sufficient to capture this semantic information, as they only consider the associated relation of chemical compounds. To capture this equipment-related relationship, "Contained" is applied.

"Contained" is defined as the association that chemical compounds are placed inside the related equipment. The direction is from the related equipment to the previous chemical compound. As in example (66), ⌈**a Teflon® flask**⌋$_3$ is used as reaction equipment and the process to this equipment is equal to the process for the reaction products inside (i.e. ⌈**Sulfuric acid (conc, 1ml)**⌋$_1$, ⌈**a solution of D-(R)-4-hydroxyphenylglycine (1.0g, 6.0mmol) in 1,4-dioxane (8ml)**⌋$_2$ and ⌈**isobutylene (8g, 142.6mmol, condensed at -78C)**⌋$_5$). This guideline annotates the relationship between these products and their related equipment as "Contained".

(66) ⌈**Sulfuric acid (conc, 1ml)**⌋$_1$ was added to ⌈**a solution of D-(R)-4-hydroxyphenylglycine (1.0g, 6.0mmol) in 1,4-dioxane (8ml)**⌋$_2$ placed in ⌈**a Teflon® flask**⌋$_3$. ⌈**The flask**⌋$_4$ was cooled to -78C and ⌈**isobutylene (8g, 142.6mmol, condensed at -78C)**⌋$_5$ was added. ⌈**The flask**⌋$_6$ was placed in an autoclave at room temperature and stirred for 15h. The autoclave was cooled on ice before opened.

- Contained (⌈**Sulfuric acid (conc, 1ml)**⌋$_1$, ⌈**a Teflon® flask**⌋$_3$)
- Contained (⌈**a solution of D-(R)-4-hydroxyphenylglycine (1.0g, 6.0mmol) in 1,4-dioxane (8ml)**⌋$_2$, ⌈**a Teflon® flask**⌋$_3$)
- Contained (⌈**isobutylene (8g, 142.6mmol, condensed at -78C)**⌋$_5$, ⌈**the flask**⌋$_6$)

(67) ⌈**Pyrazinecarboxylic acid (152.8 mg, 1.23 mmol, 1 eq)**⌋$_1$ and ⌈**H-Phe-OtBu-HCl (253.8 mg, 0.98 mmol, 0.8 eq)**⌋$_2$ were charged into ⌈**an eggplant flask**⌋$_3$...

- Contained (⌈**Pyrazinecarboxylic acid (152.8 mg, 1.23 mmol, 1 eq)**⌋$_1$, ⌈**an eggplant flask**⌋$_3$)
- Contained (⌈**H-Phe-OtBu-HCl (253.8 mg, 0.98 mmol, 0.8 eq)**⌋$_1$, ⌈**an eggplant flask**⌋$_3$)

One related issue is that "Contained" does not consider the relationship between the equipment. As in example (66), *The flask* is placed in *an autoclave* while it is not annotated as "Contrained".

In some cases (e.g. example (68) - (70)), the equipment can be semantically referring to a chemical compound. It is called zeugma, which is defined as a single phrase or word that joins different parts of a sentence. Under this observation, we can allow equipment links to chemical compounds via other bridging relations (i.e. "transformed", "reaction-associated" and "work-up") if equipment can be interpreted as a chemical compound.

(68) ⌈**6-(2-Amino-6-morpholinopyrimidin-4-yl)-3'-fluoro-5-methoxy-[2,4'-bipyridin]-2'-amine (120 mg, 301.95 μmol)**⌋$_1$ and ⌈**pyridine hydrochloride (523.41 mg, 4.53 mmol)**⌋$_2$ were stirred in ⌈**a sealed tube**⌋$_3$ at 170 ^{o}C for 30 min. ⌈**The resulting mixture**⌋$_4$ was cooled to room temperature...

- Contained (⌈**6-(2-Amino-6-morpholinopyrimidin-4-yl)-3'-fluoro-5-methoxy-[2,4'-bipyridin]-2'-amine (120 mg, 301.95 μmol)**⌋$_1$, ⌈**a sealed tube**⌋$_3$)
- Contained (⌈**pyridine hydrochloride (523.41 mg, 4.53 mmol)**⌋$_2$, ⌈**a sealed tube**⌋$_3$)
- Transformed (⌈**a sealed tube**⌋$_3$, ⌈**The resulting mixture**⌋$_4$)

(69) To ⌈**a microwave vial**⌋$_1$ charged with ⌈**EtOH (3 mL)**⌋$_2$ were added ⌈**3-bromo-4-methylpyridin-2-amine (100 mg, 0.54 mmol)**⌋$_3$, ⌈**1-bromo-3,3-dimethylbutan-2-one (0.16 mL, 1.18 mmol)**⌋$_4$, and ⌈**potassium phosphate (340 mg, 1.60 mmol)**⌋$_5$. ⌈**The vial**⌋$_6$ was capped and heated to 160 ^{o}C. in the microwave for 60 minutes. ⌈**The reaction mixture**⌋$_7$ was diluted with H2O and extracted with EtOAc (x2)...

- Coreference (⌈**a microwave vial**⌋$_1$, ⌈**The vial**⌋$_6$)
- Contained (⌈**EtOH (3 mL)**⌋$_2$, ⌈**The vial**⌋$_6$)
- Contained (⌈**3-bromo-4-methylpyridin-2-amine (100 mg, 0.54 mmol)**⌋$_3$, ⌈**The vial**⌋$_6$)
- Contained (⌈**1-bromo-3,3-dimethylbutan-2-one (0.16 mL, 1.18 mmol)**⌋$_4$, ⌈**The vial**⌋$_6$)
- Contained (⌈**potassium phosphate (340 mg, 1.60 mmol)**⌋$_5$, ⌈**The vial**⌋$_6$)
- Transformed (⌈**The vial**⌋$_6$, ⌈**The reaction mixture**⌋$_7$)

(70) To ⌈**a suspension of 6-(2-aminothiazol-4-yl)-3,4-dihydroquinolin-2(1H)-one (0.100 g, 0.408 mmol) and 4-methylthiazole-5-carboxylic acid (0.064 g, 0.448 mmol), and pyridine (0.15 mL, 1.832 mmol) in acetonitrile (4 mL)**⌋$_1$ in ⌈**a sealed tube**⌋$_2$ was added ⌈**propylphosphonic anhydride solution (50 wt % in ethyl acetate, 0.85 mL, 1.432 mmol)**⌋$_3$. ⌈**The sealed tube**⌋$_4$ was heated to 50 ^{o}C. for 4 days and ⌈**the precipitation**⌋$_5$ formed...

- Contained (⌈**a suspension of 6-(2-aminothiazol-4-yl)-3,4-dihydroquinolin-2(1H)-one (0.100 g, 0.408 mmol) and 4-methylthiazole-5-carboxylic acid (0.064 g, 0.448 mmol), and pyridine (0.15 mL, 1.832 mmol) in acetonitrile (4 mL)**⌋$_1$, ⌈**a sealed tube**⌋$_2$)
- Coreference (⌈**a sealed tube**⌋$_2$, ⌈**The sealed tube**⌋$_4$)
- Contained (⌈**propylphosphonic anhydride solution (50 wt % in ethyl acetate, 0.85 mL, 1.432 mmol)**⌋$_3$, ⌈**The sealed tube**⌋$_4$)

- Transformed (⌈**The sealed tube**⌋$_4$, ⌈**the precipitation**⌋$_5$)

Introducing zeugma can help to capture referring information in the chemical patents while also brings ambiguous to understanding if equipment can be interpreted as chemical compounds. In order to eliminate the redundancy of referring relation annotation, we also follow the restrictions:

1. For "coreference" and "contained" relationships, the equipment is the equipment itself;
2. For "transformed", "reaction-associated" and "work-up" relationships, when the equipment can be semantically interpreted as a chemical compound, it can be involved in these relationships;
3. For equipment-related bridging annotation, "contained" relationship has the priority comparing to the other bridging relationships (i.e. "transformed", "reaction-associated", "work-up")

5 Special Issues

5.1 Conflict in plural expressions

For this kind of conflict, which means the disagreement in grammatical number form of mentions, whether mentions are anaphora associated is evaluated by considering the context instead of the grammatical number.

For coreference, if there is a disagreement in number, yet it is clear that they refer to the same entity, it is acceptable to ignore the number conflict. As in example (71), ⌈**the starting material**⌋$_3$ is singular while it should be treated as a plural mention and coreferential link to the previous mentions based on the context.

(71) ⌈**A solution of 1.5 equiv. of methyliodide**⌋$_1$ in acetone was added dropwise to ⌈**an ice-cooled mixture of 1.0 equiv. of N-[4-(benzoxazol-2-yl)-methoxyphenyl]-N'-phenylthiourea and 1.0 equiv. of potassium carbonate in acetone (ca5 mL/mmol)**⌋$_2$. After stirring for 2 h at 0 ^{o}C, the mixture was warmed to room temperature and stirred at this temperature until ⌈**the starting material**⌋$_3$ was completely consumed (tlc monitoring).

- Coreference (⌈**A solution of 1.5 equiv. of methyliodide**⌋, ⌈**the starting material**⌋$_3$)
- Coreference (⌈**an ice-cooled mixture of 1.0 equiv. of N-[4-(benzoxazol-2-yl)-methoxyphenyl]-N'-phenylthiourea and 1.0 equiv. of potassium carbonate in acetone (ca5 mL/mmol)**⌋$_2$, ⌈**the starting material**⌋$_3$)

What's more, for bridging, this annotation guideline aims to capture the associated relations among the mentions. Number conflict can be neglected if the relation of mentions suits the bridging definition used in this guideline. In example (72), although ⌈**they**⌋$_5$ and ⌈**it**⌋$_6$ belong to different grammatical number, these two mentions can still be considered as "Transformed" based on the definition as they have the same components but only in different states.

(72) After ⌈**1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDCl, 860 mg, 4.5 mmol)**⌋$_1$, ⌈**1-hydroxybenzotriazole (HOBt; 860 mg, 4.5 mmol)**⌋$_2$, ⌈**diisopropylethylamine (1.6 mL, 9 mmol)**⌋$_3$ were added to ⌈**the reaction solution**⌋$_4$, ⌈**they**⌋$_5$ were stirred at room temperature for 6 hours. After the reaction was completed, ⌈**it**⌋$_6$ was extracted with ethyl acetate and water.

- Coreference (⌈**1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDCl, 860 mg, 4.5 mmol)**⌋$_1$, ⌈**they**⌋$_5$)
- Coreference (⌈**1-hydroxybenzotriazole (HOBt; 860 mg, 4.5 mmol)**⌋$_2$, ⌈**they**⌋$_5$)
- Coreference (⌈**diisopropylethylamine (1.6 mL, 9 mmol)**⌋$_3$, ⌈**they**⌋$_5$)
- Coreference (⌈**the reaction solution**⌋$_4$, ⌈**they**⌋$_5$)
- Transformed (⌈**they**⌋$_5$, ⌈**it**⌋$_6$)

6 Annotation Instrument

Based on the referring phenomena discussed above, the brat rapid annotation tool[1] is used for annotation. Annotators will be provided with the chemical patents without any annotations, as shown in figure 1, and this annotation task is to label the expressions which are involved in the referring relationship and then links them based on the referring relationship.

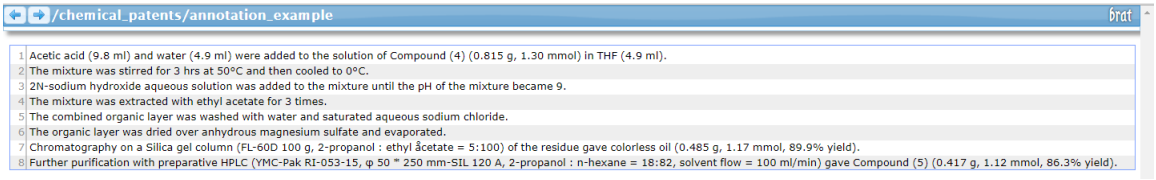


Figure 1: Example of chemical patents

To annotate the anaphora phenomena (i.e. "Coreference", "Transformed", "Reaction-associated", "Work-up" and "Contained") in the chemical patents, annotators follows the following steps. Firstly, annotators need to consider what's the referring relationships inside the corpus and then label the expressions (mentions) that are involved in these referring relationship. Furthermore, the labeled mentions are linked based on the referring relationships they are related.

[1] `https://brat.nlplab.org/`

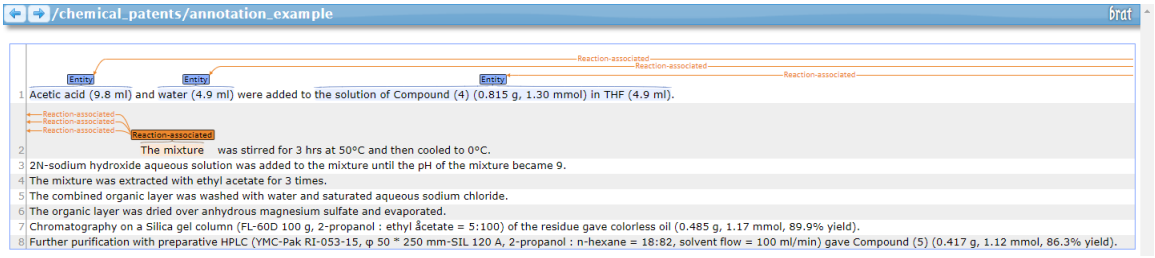
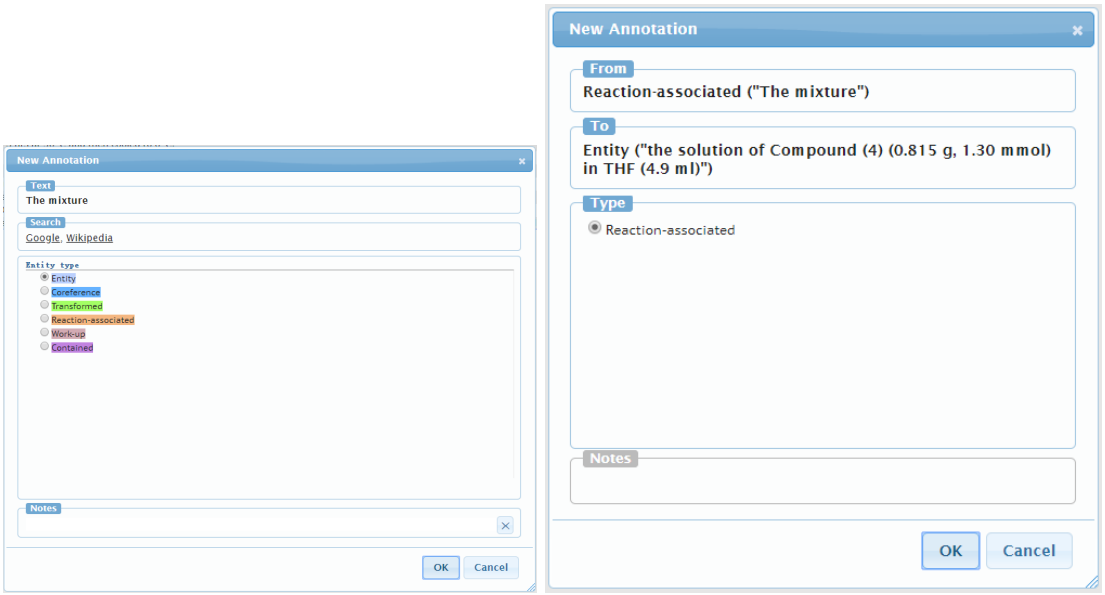


Figure 2: Annotated example of "Reaction-associated" referring relationship in the chemical patents



(a) Mention labels (b) Annotating "Reaction-associated" referring link

Figure 3: Annotation for mentions and referring links

Taking "Reaction-associated" relationship as an example. Figure 2 demonstrates how to annotate it in the chemical patent illustrated in figure 1. First of all, annotators need to think about whether there is the "Reaction-associated" relationship in the sentence. Secondly, the expressions (i.e. ⌈**Acetic acid (9.8 ml)**⌋, ⌈**water (4.9 ml)**⌋, ⌈**the solution of Compound (4) (0.815 g, 1.30 mmol) in THF (4.9 ml)**⌋, ⌈**The mixture**⌋) are labeled for "Reaction-associated" relationship and the label of mentions, as shown in figure 3(a), is annotated based on the referring relationships. Specifically, the mentions served as *antecedent* are labeled as "Entity" (e.g. ⌈**Acetic acid (9.8 ml)**⌋, ⌈**water (4.9 ml)**⌋ and ⌈**the solution of Compound (4) (0.815 g, 1.30 mmol) in THF (4.9 ml)**⌋ are *antecedent* which are labeled as "Entity" in this "Reaction-associated" referring relationship) and the mentions served as *anaphor* are annotated based on referring relationships that they are involved (e.g. ⌈**The mixture**⌋ is *anaphor*

labeled as “Reaction-associated” for “Reaction-associated” referring relationship). Lastly, based on the labels, annotators links these mentions from *anaphor* to *antecedent*. To simplify the annotation task, the referring relationships only can be linked between “Entity” (*antecedent*) and the corresponding mention label (*anaphor*). For instance, as shown in figure 3(b), “Reaction-associated” referring relationship only can be linked between “Entity” and “Reaction-associated” labels and cannot be linked with other labels.

Figure 4 shows the annotation of anaphora phenomena in the chemical patents illustrated in figure 1.

Figure 4: Annotated example of anaphora phenomena in the chemical patents

References

Saber A Akhondi, Hinnerk Rey, Markus Schwörer, Michael Maier, John Toomey, Heike Nau, Gabriele Ilchmann, Mark Sheehan, Matthias Irmer, Claudia Bobach, et al. 2019. Automatic identification of relevant chemical compounds from patents. *Database*, 2019.

Miji Choi, Justin Zobel, and Karin Verspoor. 2016. A categorical analysis of coreference resolution errors in biomedical texts. *Journal of biomedical informatics*, 60:309–318.

P Gwynne and G Heabrer. 2015. Recent developments in drug discovery: Improvements in efficiency. *Science*.

Chen Li, Maria Liakata, and Dietrich Rebholz-Schuhmann. 2013. Biological network extraction from scientific literature: state of the art and challenges. *Briefings in bioinformatics*, 15(5):856–877.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Marwin HS Segler, Mike Preuss, and Mark P Waller. 2018. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604.

Jorge A Vanegas, Sérgio Matos, Fabio González, and José L Oliveira. 2015. An overview of biomolecular event extraction from scientific documents. *Computational and mathematical methods in medicine*, 2015.