Campo	Descrição do Campo	Dataset
1. Detalhes		
Nome		ChEMU Chemical Reaction Corpus
Descrição	Deve conter uma descrição detalhada do dataset	Um corpus para pesquisa em processamento de linguagem natural (NLP) em patentes químicas, focado na extração de informações como entidades nomeadas e etapas de eventos a partir de descrições de reações químicas.
Proprietário	Criador ou Responsável pelo Dataset	Karin Verspoor, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Jiayuan He, Zenan Zhai
Versão	Identificador da versão	Versão 2
Data de criação		12 de setembro de 2020
Data da última atualização		12 de setembro de 2020
URL	-	`https://data.mendeley.com/datasets/wy6745bjfj/2`
Requisitos de Acesso	Requisitos de acesso, contato do responsável	Disponível publicamente no Mendeley Data; sem requisitos especiais de acesso.
Licença de uso	GPL, Creative Commons, etc.	Não especificada; provavelmente Creative Commons com base no padrão do Mendeley Data.
2. Características		
Tipo	Lista: Estruturado, Texto, Imagem, Texto-Imagem, Outros	Estruturado
Domínio	Descrever a área de domínio do corpus	Patentes químicas
Volume	Tamanho total (em bytes)	Aproximadamente 3MB; contém 1.500 trechos com 252.459 palavras.
Documentos	Descrição da composição. Lista: TXT, PDF, IMAGEM, JSON, outros	TXT (formato BRAT standoff: arquivos .txt e .ann)
Dados quantitativos	Descrição quantitativa e qualitativa da composição. Lista: quantidade de documentos, sentenças, tokens, ...	1.500 trechos, 7.402 sentenças, 252.459 palavras, 26.857 entidades nomeadas, 11.236 palavras gatilho, 23.445 relações.
Sanidade	Descrição de dados faltantes, com ruído, incompletos, etc	Alta qualidade; anotado manualmente por especialistas com forte concordância entre anotadores (F1 scores: 0.9760 para entidades, 0.9506 para eventos).
Método de extração	Descrição do processo (roteiro, scripts, etc)	Trechos amostrados de descrições de reações químicas em patentes, pré-identificados por especialistas; anotados manualmente por três especialistas químicos.
3. Anotações		
Tarefa	Informa a tarefa para a qual a anotação foi realizada. Classificação, Reconhecimento de Entidades Nomeadas (NER), ...	Named Entity Recognition (NER), Event Extraction (EE)

Campo	Descrição do Campo	Dataset
Classes	Descreve as classes anotadas. Deve incluir o label das classes, o significado de cada label e a distribuição delas no dataset	Etiquetas de entidades: STARTING_MATERIAL, REAGENT_CATALYST, REACTION_PRODUCT, SOLVENT, OTHER_COMPOUND, TIME, TEMPERATURE, YIELD_PERCENT, YIELD_OTHER, EXAMPLE_LABEL. Etiquetas de gatilho: REACTION_STEP, WORKUP. Etiquetas de relação: Arg1, ArgM. Distribuição fornecida no artigo.
Balanceamento	Classes distribuídas de forma igualitária em termos de amostras (binário: SIM/NÃO)	NÃO
Tipo	Deve indicar a forma como o dado foi anotado: manual, automático ou semi-automático	Manual
Método	Quando for do tipo Manual deve mencionar a quantidade de anotadores e as suas expertises. Deve descrever ou ter um link para o Guideline utilizado no processo de anotação manual	Três especialistas químicos: dois anotaram independentemente, um arbitrou discordâncias. Diretrizes de anotação disponíveis no artigo e nos arquivos suporte do conjunto de dados.
Qualidade	Indica a qualidade da anotação. No caso de anotação manual, deve informar a métrica usada para medir a concordância dos anotadores (Ex: Kappa)	Concordância entre anotadores: F1 scores de 0.9760 para entidades e 0.9506 para eventos; Cohen's Kappa também reportado.
Idioma	Informar o idioma. Valor default desse campo: Português	Inglês
Formato	Descreve o formato usado nos arquivos do dataset: CSV, IOB, 1082, XML, JSON, ...	BRAT standoff format (.txt e .ann files)
Outras informações	Mencionar aqui outras informações relevantes, tais como técnicas usadas para melhorar desbalanceamento (Ex: Data Augmentation)	Dataset dividido em conjuntos de treinamento (60%), desenvolvimento (15%) e teste (25%); verificações de similaridade para etiquetas de entidades e IPCs.
Material de Referência	Indicar documentos nos quais pode-se obter ainda mais informações, tais como: publicações, sites, ...	Artigo: "ChEMU 2020: Natural Language Processing Methods Are Effective for Information Extraction From Chemical Patents" por He et al., 2021.
Local	Para o caso dele não estar armazenado na plataforma Sinapses. Neste caso, indicar o link aqui	`https://data.mendeley.com/datasets/wy6745bjfj/2`
4. Observações		
Aspectos éticos/legais		Derivado de patentes disponíveis publicamente; sem informações pessoais ou sensíveis.
Outras observações		Focado em patentes de texto completo, oferecendo anotações mais ricas do que datasets voltados para resumos.