Guideline para Classificação de Trabalhos em PLN, Fala e Linguística Computacional

Introdução

Este documento descreve diretrizes para a classificação de trabalhos nas áreas de Processamento de Língua Natural (PLN), Tecnologias de Fala, Linguística Computacional e áreas correlatas. As categorias[1] foram definidas para garantir consistência, evitar sobreposições indevidas e estabelecer regras claras para casos especiais, como trabalhos teóricos, *surveys* e tradução. Este guia serve como base metodológica para o mapeamento do ecossistema de pesquisa apresentado no artigo **"The PROPOR Ecosystem: Structure, Roles, and Evolution of Portuguese-Language NLP"**.

1 Natural Language Processing Tasks

Classifica trabalhos cujo foco principal são **tarefas textuais específicas** ou **métodos voltados diretamente a tais tarefas**. Inclui também formulações matemáticas, análises formais, provas teóricas e estudos computacionais que tratem de tarefas de PLN. Aplicações em domínios específicos entram nessa categoria. O escopo é exclusivamente textual, não envolvendo fala. Não deve ser usado simultaneamente com a categoria de Information Extraction and Information Retrieval.

2 Natural Language Processing Applications

Abrange trabalhos que utilizam modelos, técnicas ou pipelines existentes para desenvolver **sistemas, ferramentas ou aplicações voltadas a usuários**. Sempre assume entrada textual. Ferramentas aplicadas exclusivamente a texto pertencem aqui.

[1] Classes baseadas nas diretrizes oficiais da PROPOR: `http://www.wikicfp.com/cfp/servlet/event.showcfp?copyownerid=90704&eventid=190288`

3 Natural Language Generation

Inclui qualquer tipo de método ou sistema dedicado à **geração automática de texto**, independentemente da abordagem (regras, modelos neurais, modelos generativos, etc.).

4 Information Extraction and Information Retrieval

Categoria destinada a trabalhos cujo escopo principal é claramente uma tarefa de **Extração de Informação (IE)** ou **Recuperação de Informação (IR)**. Inclui tarefas como NER, extração de relações, extração de eventos, busca, indexação e reranking. Pode-se marcar junto com *NLP tasks* ou *applications*, visto que podemos ter uma tarefa de IR ou uma aplicação de IR.

5 Speech Technologies

Equivalente à categoria *NLP Tasks*, porém voltada para **fala**. Inclui reconhecimento de fala, identificação ou verificação de locutor, compreensão de fala e síntese de fala. Engloba métodos, algoritmos e tarefas fundamentais envolvendo sinais de fala.

6 Speech Applications

Equivalente à categoria *NLP Applications*, porém dedicada à **fala**. Abrange sistemas completos baseados em tecnologias de fala, incluindo sistemas de diálogo **por voz**, tradutores fala–fala e ferramentas que integram componentes como ASR, TTS ou SLU. Sistemas de diálogo exclusivamente textuais não entram aqui.

7 Resources, Standardization and Evaluation

Inclui trabalhos que propõem ou documentam **corpora, ontologias, léxicos, treebanks, thesauri**, bem como recursos estruturados em geral. Também abrange benchmarks, campanhas de avaliação e estudos comparativos cujo foco principal é a **avaliação em escala** ou a padronização. Disponibilização do modelo não entra aqui.

8 NLP-oriented Linguistic Description or Theoretical Analysis

Classifica análises linguísticas relevantes para PLN ou descrições qualitativas e formais de fenômenos linguísticos, incluindo estudos sobre o português, suas estruturas formais (como POS e dependências como meio de análise) e análises teóricas não computacionais. A ênfase nesta categoria é **linguística** e não computacional.

9 Distributional Semantics and Language Modeling

Abrange trabalhos sobre **semântica distribucional**, criação ou análise de **embeddings**, estudos e desenvolvimento de **modelos de linguagem**, tradicionais ou neurais. Inclui análises e propostas relacionadas a representações distribucionais e a modelos de linguagem.

10 Portuguese Language Varieties and Dialect Processing

Classifica trabalhos que investigam ou comparam **variedades e dialetos da língua portuguesa**, incluindo os utilizados em Portugal, Brasil, Angola, Moçambique, Cabo Verde, Guiné-Bissau, São Tomé e Príncipe, Macau, Timor-Leste e Galícia. O foco deve ser contrastivo, comparativo ou descritivo entre variedades.

11 Multilingual Studies, Methods, Applications and Resources Including Portuguese/Galician

Categoria destinada a trabalhos genuinamente **multilíngues**, isto é, métodos, técnicas, aplicações ou recursos que operam simultaneamente em duas ou mais línguas. Não é considerado multilíngue quando o método é aplicado separadamente a diferentes idiomas sem integração. Tradução automática texto–texto pertence a esta categoria (por ser multilíngue, mas não necessariamente é exclusiva desta classe).

12 Regras Especiais

12.1 Trabalhos Teóricos

- Teoria linguística ou análises qualitativas da língua devem ser classificadas em **NLP-oriented Linguistic Description**.
- Teoria computacional sobre tarefas textuais (provas, limites, análises matemáticas) pertence a **NLP Tasks**.
- Teoria computacional aplicada à fala pertence a **Speech Technologies**.

12.2 Surveys

Surveys não constituem uma categoria própria. Devem ser classificados de acordo com o **tema central revisado**. Por exemplo, surveys sobre IE entram em IE/IR; surveys sobre modelos de linguagem entram em Distributional Semantics and Language Modeling.