

AI Instruction: Nlu Training Guide

FATAPLUS - Plateforme Agricole Madagascar

Dernière mise à jour : 10/07/2025

FATAPLUS NLU Training Guide

Introduction au NLU (Natural Language Understanding)

Le NLU permet à FATAPLUS AI de comprendre les intentions des agriculteurs, extraire les entités importantes et adapter ses réponses selon le contexte conversationnel.

Objectifs NLU FATAPLUS

1. **Classification d'Intentions** : Identifier le type de demande (conseil, produit, formation, etc.)
2. **Extraction d'Entités** : Reconnaître cultures, régions, saisons, problèmes
3. **Analyse de Sentiment** : Détecter urgence, satisfaction, frustration
4. **Gestion Contextuelle** : Maintenir le contexte conversationnel

Architecture NLU

Texte Utilisateur → Preprocessing → Intent Classification → Entity Extraction → Context Management → Structured Output

Classification d'Intentions

Intentions Principales

```
intentions_agricoles:
  conseil_technique:
    description: "Demande de conseil agricole"
    exemples:
      - "Comment planter du riz ?"
      - "Quand semer la vanille ?"
      - "Problème avec mes tomates"

  recherche_produit:
    description: "Recherche de produits/outils"
    exemples:
      - "Besoin de semences de maïs"
      - "Quel engrais pour le café ?"
      - "Outils pour irrigation"

  formation_cours:
    description: "Demande de formation"
    exemples:
      - "Cours sur l'agriculture bio"
      - "Formation élevage poulets"
      - "Apprendre la permaculture"
```

```

probleme_maladie:
  description: "Problème phytosanitaire"
  exemples:
    - "Mes plants jaunissent"
    - "Insectes sur les feuilles"
    - "Maladie du riz"

information_marche:
  description: "Informations commerciales"
  exemples:
    - "Prix du girofle"
    - "Où vendre ma récolte ?"
    - "Demande pour la vanille"

meteo_saison:
  description: "Informations météo/saisonnières"
  exemples:
    - "Quand planter avant les pluies ?"
    - "Saison sèche à Antananarivo"
    - "Préparer pour cyclone"

```

Intentions Secondaires

```

intentions_support:
  salutation:
    exemples: ["Bonjour", "Salut", "Bonsoir"]

  remerciement:
    exemples: ["Merci", "Merci beaucoup", "Misaotra"]

  demande_aide:
    exemples: ["Aidez-moi", "Je ne comprends pas", "Expliquez"]

  feedback:
    exemples: ["Très utile", "Ça marche", "Pas satisfait"]

```

Extraction d'Entités

Entités Agricoles

```

entites_principales:
  cultures:
    type: "liste_fermee"
    valeurs:
      cereales: ["riz", "maïs", "blé", "orge"]
      legumineuses: ["haricot", "pois", "lentille", "arachide"]
      tubercules: ["manioc", "patate", "pomme_terre", "igname"]
      epices: ["vanille", "girofle", "cannelle", "poivre"]
      fruits: ["litchi", "mangue", "banane", "ananas"]
      legumes: ["tomate", "oignon", "carotte", "salade"]

  regions:
    type: "liste_fermee"
    valeurs:
      provinces:
        [
          "Antananarivo",
          "Fianarantsoa",
          "Toamasina",
          "Mahajanga",
          "Toliara",
          "Antsiranana",
        ]

```

```

    villes: ["Antsirabe", "Sambava", "Morondava", "Tamatave"]

saisons:
  type: "liste_fermee"
  valeurs: ["saison_seche", "saison_pluies", "inter_saison"]

problemes:
  type: "expression_reguliere"
  patterns:
    maladies: ["maladie", "champignon", "virus", "bacterie"]
    ravageurs: ["insecte", "criquet", "chenille", "puceron"]
    climatiques: ["secheresse", "inondation", "grele", "cyclone"]

quantites:
  type: "numerique"
  patterns: ["\\d+\\s*(kg|tonne|hectare|m2|litre)"]

dates:
  type: "temporel"
  patterns: ["mois", "semaine", "janvier|fevrier|...", "demain|hier"]

```

Entités Contextuelles

```

entites_contexte:
  niveau_experience:
    valeurs: ["debutant", "intermediaire", "expert", "nouveau"]

urgence:
  indicateurs: ["urgent", "rapidement", "maintenant", "vite"]

budget:
  indicateurs: ["pas_cher", "economique", "budget_limite", "gratuit"]

preferences:
  bio: ["bio", "biologique", "naturel", "sans_chimique"]
  moderne: ["moderne", "technologique", "innovant", "nouveau"]

```

Modèles d'Entraînement

1. Classification d'Intentions

```

# Configuration modèle classification
intent_model_config = {
    "architecture": "transformer",
    "model_base": "xlm-roberta-base", # Support multilingue
    "fine_tuning": {
        "learning_rate": 2e-5,
        "batch_size": 16,
        "epochs": 5,
        "warmup_steps": 100
    },
    "data_augmentation": {
        "paraphrasing": True,
        "back_translation": ["en", "mg"], # Français ↔ Anglais ↔ Malagasy
        "synonym_replacement": True
    }
}

```

2. Extraction d'Entités (NER)

```

# Configuration NER
ner_model_config = {

```

```

"architecture": "bert_crf",
"model_base": "xlm-roberta-base",
"labels": ["B-CULTURE", "I-CULTURE", "B-REGION", "I-REGION",
           "B-PROBLEM", "I-PROBLEM", "B-QUANTITY", "I-QUANTITY"],
"training": {
  "learning_rate": 3e-5,
  "batch_size": 32,
  "epochs": 10,
  "dropout": 0.1
}
}

```



Données d'Entraînement

1. Collecte de Données

```

# Sources de données pour entraînement
data_sources = {
  "conversations_existantes": {
    "source": "logs_fataplus",
    "volume": "10000_conversations",
    "qualite": "reel_utilisateur"
  },
  "donnees_synthetiques": {
    "generation": "templates_variations",
    "volume": "50000_exemples",
    "qualite": "controlee"
  },
  "experts_agricoles": {
    "source": "interviews_specialistes",
    "volume": "2000_qa_pairs",
    "qualite": "tres_haute"
  },
  "forums_agricoles": {
    "source": "web_scraping",
    "volume": "20000_posts",
    "qualite": "variable"
  }
}

```

2. Annotation des Données

```

# Schéma d'annotation
annotation_schema = {
  "format": "IOB2", # Inside-Outside-Begin
  "outils": ["Label Studio", "Prodigy", "Doccano"],
  "guidelines": {
    "cultures": "Annoter nom complet (ex: 'riz rouge' pas juste 'riz')",
    "regions": "Inclure villages/districts si mentionnés",
    "problemes": "Capturer symptômes et causes",
    "quantites": "Inclure unité de mesure"
  },
  "qualite": {
    "inter_annotator_agreement": "> 0.85",
    "double_annotation": "20% des données",
    "expert_validation": "100% échantillon test"
  }
}

```



Pipeline d'Entraînement

1. Preprocessing

```
# Étapes de préparation
preprocessing_steps = {
    "nettoyage": {
        "normalisation_unicode": True,
        "suppression_caracteres_speciaux": False, # Garder accents
        "correction_orthographe": True,
        "detection_langue": True
    },
    "tokenisation": {
        "tokenizer": "xlm-roberta-tokenizer",
        "max_length": 512,
        "padding": "max_length",
        "truncation": True
    },
    "augmentation": {
        "paraphrase_generation": True,
        "noise_injection": 0.1,
        "entity_replacement": True
    }
}
```

2. Entraînement Multi-tâches

```
# Configuration multi-task learning
multitask_config = {
    "taches": {
        "intent_classification": {
            "poids": 0.4,
            "metrique": "f1_weighted"
        },
        "entity_extraction": {
            "poids": 0.4,
            "metrique": "f1_entity"
        },
        "sentiment_analysis": {
            "poids": 0.2,
            "metrique": "accuracy"
        }
    },
    "partage_representations": True,
    "optimisation_jointe": True
}
```

Adaptation Multilingue

1. Support Français-Malagasy

```
# Configuration multilingue
multilingual_config = {
    "langues_principales": ["fr", "mg"],
    "detection_langue": {
        "automatique": True,
        "fallback": "fr",
        "confiance_minimum": 0.7
    },
    "traduction": {
        "service": "google_translate_api",
        "cache": True,
        "post_processing": "correction_agricole"
    },
    "vocabulaire_mixte": {
```

```

    "termes_malagasy": "conserver_original",
    "expressions_locales": "annoter_traduction",
    "code_switching": "gerer_automatiquement"
  }
}

```

2. Dictionnaire Agricole Bilingue

```

# Dictionnaire spécialisé
agricultural_dictionary = {
  "cultures": {
    "vary": {"fr": "riz", "context": "cereale_base"},
    "katsaka": {"fr": "maïs", "context": "cereale_secondaire"},
    "mangahazo": {"fr": "manioc", "context": "tubercule"},
    "voatavo": {"fr": "courge", "context": "legume_fruit"}
  },
  "techniques": {
    "fambolena": {"fr": "culture", "context": "technique_generale"},
    "famafazana": {"fr": "semis", "context": "technique_semis"},
    "fanondrahana": {"fr": "irrigation", "context": "technique_eau"}
  }
}

```

Évaluation et Métriques

1. Métriques par Tâche

```

evaluation_metrics = {
  "intent_classification": {
    "accuracy": "precision_globale",
    "f1_weighted": "f1_pondere_classes",
    "confusion_matrix": "analyse_erreurs",
    "per_class_metrics": "performance_par_intention"
  },
  "entity_extraction": {
    "f1_entity": "f1_niveau_entite",
    "precision_recall": "par_type_entite",
    "exact_match": "correspondance_exacte",
    "partial_match": "correspondance_partielle"
  },
  "overall_pipeline": {
    "end_to_end_accuracy": "precision_complete",
    "response_relevance": "pertinence_reponse",
    "user_satisfaction": "satisfaction_utilisateur"
  }
}

```

2. Tests A/B

```

ab_testing_nlu = {
  "scenarios": {
    "modele_base_vs_fine_tuned": {
      "baseline": "xlm-roberta-base",
      "variant": "fataplus_fine_tuned",
      "metriques": ["accuracy", "latency", "satisfaction"]
    },
    "pipeline_simple_vs_multitask": {
      "baseline": "modeles_separes",
      "variant": "multitask_joint",
      "metriques": ["f1_overall", "consistency"]
    }
  }
}

```

```

    "duree": "2_semaines",
    "trafic_split": "50/50",
    "critere_succes": "amelioration_5_percent"
}

```



Déploiement et Monitoring

1. Pipeline de Production

```

production_pipeline = {
    "preprocessing": {
        "latence_max": "50ms",
        "cache_tokenization": True
    },
    "inference": {
        "batch_processing": True,
        "gpu_acceleration": True,
        "model_quantization": "int8"
    },
    "post_processing": {
        "confidence_filtering": 0.7,
        "entity_validation": True,
        "response_formatting": True
    }
}

```

2. Monitoring Continu

```

monitoring_nlu = {
    "metriques_temps_reel": {
        "latence_inference": "< 100ms",
        "accuracy_drift": "detection_degradation",
        "distribution_intentions": "monitoring_shift",
        "nouvelles_entites": "detection_automatique"
    },
    "alertes": {
        "accuracy_drop": "< 85%",
        "high_latency": "> 200ms",
        "unknown_intents": "> 10%",
        "entity_extraction_fail": "> 15%"
    },
    "retraining_triggers": {
        "performance_degradation": True,
        "new_data_volume": "1000_new_samples",
        "domain_shift": True
    }
}

```



Amélioration Continue

1. Active Learning

```

active_learning_config = {
    "strategies": {
        "uncertainty_sampling": "echantillons_incertain",
        "diversity_sampling": "echantillons_diversifies",
        "query_by_committee": "desaccord_modeles"
    },
    "annotation_budget": "100_samples_per_week",
    "retraining_frequency": "monthly",
}

```

```
"human_in_the_loop": True
}
```

2. Feedback Loop

```
feedback_integration = {
  "sources": {
    "user_corrections": "corrections_utilisateur",
    "expert_validation": "validation_experts",
    "conversation_outcomes": "succes_conversations"
  },
  "processing": {
    "confidence_weighting": True,
    "consensus_requirement": "2_experts_minimum",
    "automated_validation": "rules_based"
  },
  "integration": {
    "training_data_update": "automatic",
    "model_retraining": "scheduled",
    "a_b_testing": "continuous"
  }
}
```

Version : 1.0

Dernière mise à jour : Décembre 2024

Prochaine révision : Mars 2025

Guide complet pour développer un système NLU de pointe pour l'assistance agricole FATAPLUS, adapté au contexte malgache et multilingue.

FATAPLUS - Plateforme Agricole Numérique Madagascar

Contact: contact@fata.plus | Web: <https://fata.plus>

Document généré le 10/07/2025 à 09:49:58